分类号_____ 密级 _____
U D C _____ 编号 _____

硕 士 学 位 论 文

论文题目 基于非负矩阵分解的函数型矩阵填充方法研究与应用

研 究 生 姓 名：马文娟

指导教师姓名、职称：高海燕、教授

学 科、专 业 名 称：统计学、数理统计学

研 究 方 向：复杂数据分析

提 交 日 期：2024 年 6 月 3 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：马文娟 签字日期：2024年6月3日

导师签名：高海燕 签字日期：2024年6月3日

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意(选择“同意”/“不同意”)以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊(光盘版)电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：马文娟 签字日期：2024年6月3日

导师签名：高海燕 签字日期：2024年6月3日

Research and Application of Functional Matrix Completion Method based on Non-negative Matrix Factorization

Candidate: Wenjuan Ma

Supervisor: Haiyan Gao

摘要

数据重构是从部分已观测数据恢复原始数据的过程，是缺失处理的核心任务。数据重构的关键在于充分、合理地利用数据的特性。目前，随着函数型数据分析的发展，函数型矩阵填充方法成为数据重构的主流方法之一。然而，现有的函数型矩阵填充方法在数据学习过程中未充分利用数据的潜在特征，如样本曲线的相关性、样本的高阶邻域信息等。为了解决这些问题，本文在函数型数据分析框架下，通过引入非负约束，借助非负矩阵分解(Non-negative Matrix Factorization，NMF)和类信息，提出融合类信息的函数型矩阵填充方法(Non-negative Functional Matrix Completion Method with Class Information，CNFMC)。同时，利用样本信息的高阶邻域关系以及各视角之间的多样性估计缺失数据，提出基于图正则化的多视角函数型矩阵填充方法(Based on Graph Regularization Multi-view Non-negative Functional Matrix Completion，GMVNFMC)。本文主要研究内容包括以下两部分：

(1) 提出一种融合类信息的函数型矩阵填充方法(CNFMC)。基于非负矩阵分解构造函数型矩阵填充方法，在此基础上通过聚类划分引入样本类信息，借助类内样本相关性插补缺失值，并采用自加权集成学习算法动态赋权重计算得最终插补值。在公共交通数据集 PeMS 进行了缺失模拟插补实验，并针对空气质量缺失数据进行实证应用分析。结果表明：相较于 K 近邻算法、MICE、PACE 等 10 种插补方法，CNFMC 方法插补精度高、鲁棒性好、适用性较强，且耗时可控，能够保证插补的有效性和准确性。

(2) 提出一种基于图正则化的多视角函数型矩阵填充方法(GMVNFMC)。通过引入最优图正则化，充分考虑了各视角内样本信息的高阶邻域关系，减少了信息损失；同时，利用希尔伯特-施密特独立性准则探索不同视角之间包含的互补信息，进而提高插补精度。分别对空气污染物数据集进行模拟插补实验和实证应用，结果表明，相较于其他主流插补方法，GMVNFMC 方法具有更好的插补效果。

关键词：函数型数据分析 矩阵填充 非负矩阵分解 多视角学习 缺失插补

Abstract

Data reconstruction is the process of restoring original data from partially observed data, and is the core task of missing data handling. The key to data reconstruction lies in fully and reasonably utilizing the characteristics of data. Currently, with the development of functional data analysis, functional matrix completion method has become one of the mainstream methods for data reconstruction. However, existing functional matrix completion methods do not fully utilize the potential features of data in the data learning process, such as the correlation of sample curves and high-order neighborhood information of samples. To address these issues, this thesis proposes a non-negative functional matrix completion method with class information (CNFMC) that by introducing non-negative constraints and utilizing non-negative matrix factorization (NMF) and class information within the framework of functional data analysis. Meanwhile, utilizing the high-order neighborhood relationship of sample information and the diversity between different views to estimate missing data, a multi-view non-negative functional matrix completion (GMVNFMC) method based on graph regularization is proposed. The main research content of this thesis includes the following two parts:

(1) Propose a non-negative functional matrix completion method with class information (CNFMC). Based on the non-negative matrix factorization constructor functional matrix completion method, sample class information is introduced through clustering partitioning, missing values are imputed using intra class sample correlation, and the final imputation value is calculated by dynamically assigning weights using a self-weighted ensemble learning algorithm. We conducted missing simulation imputation experiments on the public transportation dataset PeMS and conducted empirical application analysis on air quality missing data. The results show that compared to 10 interpolation methods such as K-nearest neighbor algorithm, MICE, PACE, etc., the CNFMC method has high interpolation accuracy, good robustness, strong applicability, and controllable time consumption, which can ensure the effectiveness and accuracy of imputation.

(2) Propose a multi-view functional matrix completion method based on graph regularization (GMVNFMC). By introducing optimal graph regularization, the high-order neighborhood relationships of sample information within each view are fully considered, reducing information loss; Meanwhile, utilizing the Hilbert-Schmidt independence criterion to explore the complementary information contained between different views,

thereby improving interpolation accuracy. Simulated interpolation experiments and empirical applications were conducted on the air pollutant datasets, and the results showed that the GMVNFMC method has better imputation performance compared to other mainstream imputation methods.

Keywords: Functional data analysis; Matrix completion; Non-negative matrix factorization; Multi-view learning; Missing imputation

目录

1 绪论

1.1 研究背景

随着物联网、互联网的飞速发展和智能终端的普及，数据正以前所未有的形式、速度、广度不断增长和累积，大数据现象深刻地影响着经济社会发展的方方面面。在各类大数据源中，物联网传感数据、智能终端监测数据成为了一种重要的数据资源，其中一些样本数据的外在表现形式虽是离散、稀疏的片段点集，但内在结构却呈现出连续、动态的函数曲线(或曲面)特征，如路网中监测器在某一时间段贮存的交通流量数据、不同地区的多期温度与降雨量数据、同一地区不同区域某一时间段的空气污染物浓度数据等。如果将这些具备函数特征的数据看成一个整体进行统计分析，则称为函数型数据分析(FDA)(Ramsay & Silverman，2005)。一般来讲，函数型数据是不能直接观测的，它是指数据的潜在生成过程为光滑的函数过程时对其进行观测所得到的离散记录。然而，在观测过程中，由于某些不可预期的因素导致离散采样点存在大量缺失。例如，在智能交通系统中，临时软件或硬件故障、维护操作、传输失真以及传输期间数据包丢失等原因使得各种电子设备和技术收集的交通流量数据往往不完整，含有缺失值；在环境监测中，实时发布的污染物浓度数据，由于设备、电源、传输以及监测点增设或停运等原因，导致监测结果存在大量的缺失；在网上拍卖中，由于拍卖出价时间分布稀疏且不规律，出价信息缺失。众所周知，数据缺失不仅导致分析结果的准确性受到影响，降低模型的性能，增加过拟合的风险，从而导致模型的泛化能力下降，还会不同程度地增大统计分析的复杂性和难度、降低统计推断的精度，最终导致统计分析结果偏误。因此，如何科学有效地修复函数型数据缺失是研究热点问题之一。

1.2 研究目的与意义

本文关注稀疏函数型数据的修复问题，当缺失率较高或者出现大规模缺失时，传统插补方法不能准确地估计缺失值，有效性降低。函数型数据修复的关键在于利用已观测数据进行数据重构，目前函数型矩阵填充是数据重构的主流方法之一。

该方法在处理函数型缺失数据时，相比其他传统插补方法可以处理大规模缺失，且具有更高的精度。然而，现有函数型矩阵填充方法在数据学习以及矩阵重构过程中主要存在以下两个问题：(1) 大多以全局信息为主，缺乏对样本信息局部特征的考虑；(2) 只考虑数据矩阵中条目之间的线性关系，而忽略潜在的非线性关系，这在一定程度上降低了重构性能。因此，本文分别针对这两个问题，通过引入不同的样本信息，探讨函数型矩阵填充方法在特定场景下如何有效地挖掘原始数据的潜在表示信息，全面地对数据进行表征，充分发挥函数型矩阵填充方法的优势，更加准确、快速地实现缺失值的估计。

从方法上，本文以函数型矩阵填充技术为核心，结合函数型聚类方法分析样本曲线的相关信息，利用流形学习技术探索样本之间的非线性关系等，充分挖掘函数型数据中的深层结构信息，进一步提高估计的准确性，为函数型缺失数据处理提供理论支持。此外，本文将所提方法应用于交通流量数据和空气质量数据来验证其有效性和适用性，结果表明，所提方法不仅具有较高的估计精度，且适用性强、运行时间短。因此，本文对函数型矩阵填充方法的研究在理论与应用上均具有一定的意义与价值。

1.3 文献综述

(1) 传统的缺失处理方法

基于数据的缺失机制，对缺失数据进行处理。传统离散数据的缺失处理方法可以大致分为三大策略：不做处理、删除法、插补法。缺失数据处理中不做处理是最简单的一种方法，适用于缺失率较小的样本，然而数据的不完整性使得统计分析方法性能较差、应用时间效率低；删除法包括简单删除法和权重删除法，删除不仅会造成数据信息的二次损失，得出错误的结论，而且删除后的数据不具有代表性，无法反映整个研究的实际情况；插补法作为缺失数据最常见且较有效的处理方法，其主要有：均值填充(Roth，1994)、线性插值(Blu & Thevenaz，2004)、热卡插补法(Bertsimas 等，2017)、K 最近邻插补法(KNN)(Rumaling 等，2020)、EM(Expectation Maximization)插补(Dempster 等，1977)、MICE(Van，2007)、贝叶斯模型等(Henry & Kyburg，1988)，其中，均值填充只适用于变量服从正态或近似正态的情况，且不能反映缺失数据的变异性；当数据高度集中时，热卡插补

法填充效果差，并在模拟数据的分布特征时缺乏准确性；涉及“距离度量”的插补方法也会导致处理效果不理想，如 K 最近邻插补法适用于离散数据，当缺失比例较大或缺失数据点大量连续时插补效果较差；EM 插补要求数据分布服从正态分布，且当数据量大时计算过程较为繁琐、算法收敛速度过于依赖初值的选择，算法效率低；多重插补考虑了缺失数据的不确定性，插补效果虽好但计算量巨大，不适用于处理大规模缺失数据(陈小波等，2019)；贝叶斯模型中极大似然估计要求模型的形式必须正确，参数形式不正确将导致错误的结论。同时，该方法适用于数据规模较小的情况，当缺失率较高时，估计参数增加，使得估计方差增大，降低填充效果；支持向量机的参数选择较为困难，当训练样本数目太大时计算速度慢(Jerez 等，2010)；基于最大似然估计(MLE)的方法中(Qu 等，2009；Shi 等，2013)，如概率主成分分析(PPCA)采用特定的参数模型，可以同时实现模型拟合和缺失数据插补。但在估计模型参数时，由于 EM 算法的固有特性，当缺失率较高时 PPCA 插补效果较差(Dempster 等，1977)。基于回归的插补方法试图构建从已知属性到缺失属性的映射函数。其中局部最小二乘(LLS)插补(Shi 等，2013)最为典型，当出现大规模缺失时，LLS 插补精度不高。上述方法大部分只适用于小规模缺失数据处理，针对这一问题，本文重点介绍另一种离散数据缺失处理方法——矩阵填充技术。

近年来，矩阵填充技术在机器学习领域得到了充分发展(Wang 等，2023；Xie 等，2023)，其目标是根据已有的观测数据对缺失数据进行预测和恢复，用以处理矩阵数据的大规模缺失问题。该问题的主要求解策略是基于核范数最小化，这需要计算奇异值分解(SVD)，随着底层矩阵的规模和秩的增加，奇异值分解的计算复杂度越来越高，而由于矩阵分解(Matrix Factorization，MF)避免了 SVD 计算，故当完备矩阵是低秩矩阵时，基于 MF 的方法是解决矩阵填充问题的常用方法之一。MF 的主要思想是将观测矩阵分解成两个(或多个)低秩潜变量。由于 MF 在潜在特征学习中的高效性，该方法被广泛用于矩阵填充问题中。例如，Koren 等(2009)最早将 MF 模型应用于推荐系统中；进一步，为提高推荐系统的性能，Han 等(2018)利用标签之间的相关性，提出了一种扩展标签诱导矩阵分解技术；Wang 等(2018)提出置信度感知矩阵分解框架(CMF)，在矩阵分解过程中为用户和项目引入方差参数，从而提高推荐结果的质量；Lara-Cabrera 等(2020)提出自动生成

矩阵因子分解的进化方法(EMF)；在求解算法方面，Wen 等(2012)提出了一种低秩分解模型，并构建 SOR 算法进行模型的求解，研究表明，该算法可以有效地解决众多问题，其速度至少是许多核范数最小化算法的几倍；Jain 等(2012)研究了矩阵填充的交替最小化算法，结果表明，交替最小化算法可以保证更快地达到收敛；Gu 等(2023)提出了鲁棒的交替最小化算法，相比其他的算法，该算法对误差的容忍度更高。与其他插补方法相比，矩阵填充技术的预测精度更高(陈小波等，2019)。然而，经典的矩阵填充把样本看作一个整体，潜在地假设所有样本数据同等重要，更加强调样本的共性，在数据重构时未能考虑到数据自身固有的特征和结构信息，当数据缺失模式复杂、多样且缺失率较高时，不可避免地降低了矩阵填充的性能，其效果往往不理想。此外，由于低秩性源于线性潜变量模型(Candes 等，2011)，在矩阵填充问题中仅提供观测数据的线性重构，即对缺失数据的估计是由潜在特征向量的线性交互得到，而当部分数据来源于非线性潜变量模型时，这一系列方法的性能下降，为此，有学者利用深度神经网络的非线性激活函数近似非线性函数提出了一系列深度矩阵填充方法。Sedhain 等(2015)提出基于自编码的协同过滤(AECF)，首次将基于自编码的方法应用于非线性矩阵填充问题中；对于推荐系统，Zhuang 等(2017)提出一种通过双自编码器推荐的表示学习框架，另一种基于自编码的方法由 Fan & Chow(2017)提出，该方法将部分观测数据与深度神经网络框架相结合；Fan & Cheng(2018)提出深度矩阵分解方法(DMF)用于解决非线性矩阵填充；Fan 等(2022)针对 IncRNA 的预测提出了基于图卷积的深度矩阵填充方法；Ye 等(2023)提出了预测地铁 OD(Origin-destination)的深度矩阵填充方法，并利用深圳和杭州地铁系统的智能卡数据证明了该方法的有效性。上述方法的初衷在于结合深度神经网络以寻找更一般的潜在因子用于精确地重构观测矩阵，但潜变量位于非线性子空间并且只学习非线性潜变量的假设，导致这些模型被限制于处理矩阵中所有元素之间仅具有非线性关系，而忽略了线性关系。然而，实际数据集的条目之间并没有完全的非线性(线性)关系。因此，尽管这些模型对观测数据的重构优于线性模型，但仍不能提供更一般的潜变量。而流行学习是一种旨在克服这些挑战的方法之一。

流形学习作为典型的非线性约简方法，假设如果任意两个样本点x_i与x_j在原始高维特征空间中是一对近邻点，那么该对近邻点在降维后的低维表示空间中也保持近邻关系(Belkin & Niyogi，2001)。因此，流形学习能将高维空间中的独有

几何拓扑关系映射至低维空间的样本点，且使得捕获各样本点在低维空间的对应关系成为可能。根据谱图理论以及流行学习理论，样本点之间的局部几何关系可以用基于欧式距离的近邻图来近似表示。基于上述理论，Cai 等(2008)提出了基于图正则化的非负矩阵分解(GNMF)算法。GNMF 算法将数据空间建模为嵌入在周围空间中的子流形，在分解过程中利用拉普拉斯图正则化项，保持了原始数据的高维几何结构，并使其在降维过程中仍然保持对应的映射关系。相较于仅考虑数据欧几里得结构的一般 NMF，GNMF 具有更高的数据表示性能和子空间学习能力。近年来，传统矩阵填充方法虽然考虑了观测数据的整体结构信息，但忽略了输入矩阵的几何结构信息，故基于图正则化的矩阵分解方法得到了广泛的推广和应用(Yi 等，2019；Jiao 等，2020；Jain 等，2023；Fang 等，2023)，然而，现有大多数方法在构造图正则化项的相似矩阵时仅考虑样本的一阶关系，不能充分挖掘样本空间中固有的几何关联深层结构信息。故在矩阵填充问题中，如何从观测矩阵中学习到更多的数据关系以及更具代表性的潜变量已成为众多学者关注的问题之一。

(2) 函数型数据缺失处理方法

利用数据的函数性特征，可以从函数的视角对缺失数据进行处理(Laird & Ware，1982)。FDA 在处理缺失数据时具有优势，它放松了数据采集的结构约束和分布设定，能够从交互的函数视角深层次挖掘数据潜在的动态信息。为此，人们有针对性地提出了一些函数型数据缺失处理技术。一种策略是忽略缺失数据，在数据分布被假定的基础上对包含缺失值的数据直接进行建模分析。这种做法属于纵向数据分析方法，常采用混合效应模型并利用 EM 算法进行处理(Laird & Ware，1982；James & Hastie，2000； James 等，2002；Peng & Debashis，2009；张淑楠，2018)。然而，混合效应模型只适用于每个样本曲线的观测值数目足够多的情形。对于观测值数目较小的情形，另一种策略是将缺失数据补齐，进而针对完整数据开展分析。对于函数型数据而言，数据的变动轨迹通常包含在前几个函数型主成分张成的子空间中，函数型主成分分析(FPCA)(Karhunen-Loève 展开，K-L 展开)是函数型数据生成的一种重要方法(Ramsay & Dalzell，1991；Yao 等，2005； Horváth & Kokoszka，2012)。Li & Chiou(2021)基于 K-L 展开、子空间投影技术以及 FPCA 提出“聚类+插补”并行的一步法，研究表明，该方法的插补性能优于其他函数型数据插补方法，并且类间信息有助于提高插补精度。

此外，函数型数据缺失处理可以转化为曲线轨迹预测问题，进而转化为矩阵填充问题(Rennie & Srebro，2005； Candes & Recht，2012)。在一定条件下，函数型数据修复问题等价于秩约束的矩阵填充问题(Descary & Panaretos，2019)，从而通过对基函数施加连续性约束，并对观测数据矩阵使用稀疏矩阵分解技术来进行处理(Kidziński & Hastie，2018)。Kidziński & Hastie 基于这一思想，提出软函数型数据填充法(SFI)和硬函数型数据填充法(HFI)，然而，当处理非负函数型数据时，如空气质量数据等，以上两种插补方法不能保证结果非负。为此，进一步对于非负数据，薛娇等(2022)在 SFI、HFI 方法的基础上，提出一种基于多视角学习的非负函数型矩阵填充(MVNFMC)方法。尽管 MVNFMC 方法具有较好的插补性能，但在数据修复过程中，MVNFMC 方法仅考虑由不同视角间共享的共公信息，而忽略了包含在不同视角中的互补信息。综上所述，本文在 Kidziński & Hastie 方法以及 MVNFMC 方法的基础上，以函数型缺失数据为研究对象，构造函数型矩阵填充方法，用以克服现有函数型数据填充方法的局限性。

(3) 多视角学习插补方法

近年来，学者们提出了许多通过考虑不同视角的多样性来从多视角数据中学习的方法。这些视角可以从多个源或不同特征子集中获得。例如，一个人可以通过人脸、指纹、签名或者虹膜来识别，即信息来自多个源，而图像可以通过其颜色或纹理特征来表示，这些特征可以被视为图像的不同特征子集。多视角学习的目标是学习不同领域或者各种特征提取器的特征。样本在不同的视角中采用不同的形式，表现出异构的属性，但同时保留了不同视角间的相似信息。多视角学习致力于利用多个视角中的互补性，自适应地学习不同视角之间的关系。充分利用多视角可以极大地提高多视角模型的性能。因此，多视角学习因其良好性能在许多应用中备受关注(Zhu 等，2015；Wang 等，2013；Liu 等，2023；Liu 等，2023)，

多视角学习在缺失数据插补领域的应用尤为广泛。Chen等(2015)利用时空K近邻对缺失文本数据进行插补；Yi等(2016)通过考虑时间序列数据的时间相关性和空间相关性，提出基于时间、空间两个视角的缺失值数据插补模型(ST-MVL)；Qin等(2019)考虑浮标数据的时间、空间以及各个变量之间的相关性，对每个视角应用不同的矩阵填充方法并利用集成学习算法对不同的估计结果进行融合，提出基于多视角学习的矩阵填充方法(MC-MVL)；Liu等(2016)基于多视角学习方法，融合来自时间、空间不同视角的数据集研究城市水质预测问题，并将所提方法应

用于多个数据集进行验证，结果表明该方法具有一定的有效性；以及张贝娜等(2019)考虑指数移动平均、普通克里金以及矩阵填充方法，提出针对缺失空气质量数据的处理方法——基于时空多视图BP神经网络的数据补全方法；Gong等(2021)对于多视角城市统计缺失数据，利用NMF提出了一种改进的时空多核插补方法(SMV-NMF)；在估计城市GPS变化轨迹时，Zhang等(2023)基于空间视角和时间视角提出了一种多视角插补方法(MVHGN)。上述研究表明，同等条件下，相比仅考虑一个维度信息的单视角缺失数据处理方法，多视角学习方法具有更高的插补精度。

1.4 研究内容与结构

本文主要基于函数型矩阵填充方法进行研究工作，论文包含五个章节，具体研究内容安排如下：

第 1 章，绪论。本章主要介绍了函数型矩阵填充方法的研究背景、目的和意义、缺失数据处理方法的相关文献概述；其次，给出了论文的结构安排和创新点。

第 2 章，基础理论与方法。包括数据缺失机制、函数型数据曲线拟合、基于非负矩阵分解的矩阵填充、最优图正则化项、希尔伯特-施密特独立性准则以及缺失插补方法评价指标。

第 3 章，融合类信息的函数型矩阵填充方法(CNFMC)。本章考虑到结合类信息可以进一步提高缺失数据的插补精度，在构建函数型矩阵填充方法时利用数据的潜在变化模式，即通过聚类划分引入样本类信息，使得每一类样本相似度高，并借助类内样本相关性插补缺失值。其次，为充分利用不同聚类中的插补结果以及降低聚类数的影响，利用集成学习方法将不同类别下每条样本的插补结果进行融合，构建基于类信息的函数型矩阵填充方法。

第 4 章，基于图正则化的多视角函数型矩阵填充方法(GMVNFMC)。为进一步提高缺失数据修复效果，在重构观测矩阵的过程中，引入最优图正则化项，不仅可以将数据矩阵条目中的非线性关系纳入到模型之中，并且通过考虑数据之间的高阶邻域关系，充分利用各个视角的特征空间关联结构信息，增强子空间整体学习能力，提高模型的鲁棒性；同时考虑了不同视角中包含的互补信息，构建基于图正则化的多视角函数型矩阵填充方法。

第 5 章，总结及展望。对论文的主要研究工作与结果进行总结，并给出了后续研究工作的方向。

本文研究技术路线如图 1.1 所示。

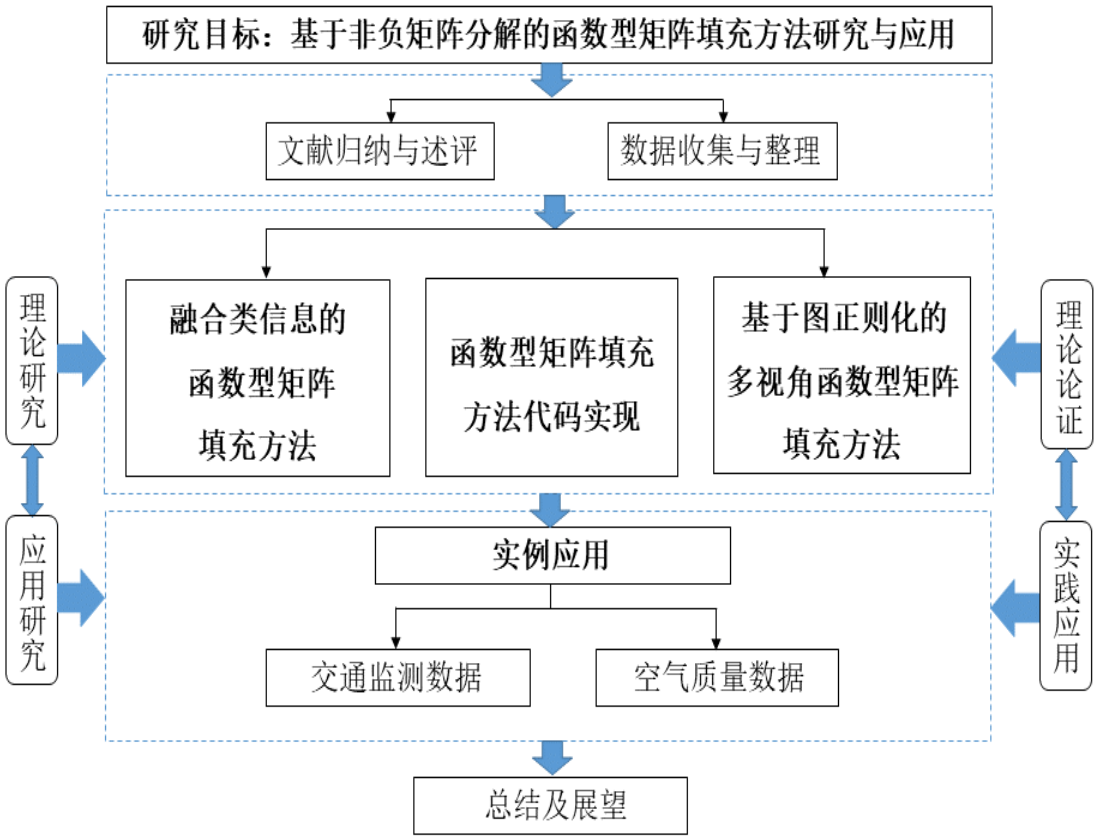


图 1.1 技术路线图

1.5 创新点

本文在函数型数据分析框架下，开展函数型矩阵填充方法研究，国内外有关该领域的研究较少。本文创新点有：

(1) 将函数型矩阵填充问题与聚类分析方法相结合，进一步提高了缺失数据修复效果；

(2) 将函数型矩阵填充方法与流形学习相结合，在 NMF 的线性子空间中学习观测数据从高维空间到低维空间的非线性映射关系，从而在数据重构过程中同时考虑矩阵条目之间的线性关系和非线性关系，降低数据重构误差，提高缺失值插补性能；

(3) 在多视角函数型矩阵填充问题中引入最优图正则化和希尔伯特-施密特独立性准则，在数据重构过程中充分利用各个视角特征空间关联结构以及更多潜在信息，增强子空间的学习能力，有助于提高函数型矩阵填充方法的插补精度。

2 基础理论与方法

2.1 数据缺失机制

对缺失数据的处理过程中，确认数据集的缺失机制是至关重要的一步，了解缺失机制有利于选择合适的算法对缺失数据集进行有效的处理。

设数据矩阵$\boldsymbol{Y}$是m组观测数据、n个观测属性构成的$m \times n$矩阵，定义$\boldsymbol{M}$为m行n列矩阵且

$$\boldsymbol{M}_{ij} = \begin{cases} 0 & \text{for } \boldsymbol{Y}_{ij} \in \boldsymbol{Y}_m \\ 1 & \text{for } \boldsymbol{Y}_{ij} \in \boldsymbol{Y}_o \end{cases}$$

其中，矩阵$\boldsymbol{M}$决定了数据矩阵$\boldsymbol{Y}$的缺失情况，根据$\boldsymbol{M}$矩阵可以将数据矩阵$\boldsymbol{Y}$分为已观测和缺失的两部分，分别记为$\boldsymbol{Y}_o$与$\boldsymbol{Y}_m$。数据缺失的机制基于似然性可以分为三种：完全随机缺失(Missing Completely At Random，MCAR)、随机缺失(Missing At Random，MAR)和非随机缺失(Not Missing At Random，NMAR)。

(1) 完全随机缺失

当数据缺失的概率与原则上应该获得的具体缺失值以及观测到的变量均无关时，称为完全随机缺失。例如，在一项调查中，某些问题由于技术故障未能记录下来；在医疗记录中，医生可能会漏填某些重要的检查结果；在金融交易中，交易员可能会故意隐瞒某些交易信息等等。完全随机缺失机制的数学表达为：

$$P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}) \tag{2.1}$$

式(2.1)表明数据的缺失既与自身的取值无关，也与完全观测、不完全观测变量无关。这种类型的缺失通常对研究结果的影响较小。

(2) 随机缺失

当数据缺失的概率取决于观测的变量，但与原则上应该获得的具体缺失值无关时，称为随机缺失。例如，在一项健康调查中，一些参与者可能选择不回答某些敏感问题，这些缺失的数据与敏感问题本身无关，但可能与参与者的年龄、性别等因素有关；在医疗记录中，医生可能会因为忙碌而忽略某些重要的检查结果；在金融交易中，交易员可能会因为情绪波动而故意遗漏某些交易

信息等等。随机缺失的数学表达式为：

$$P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}|\boldsymbol{Y}_o) \tag{2.2}$$

式(2.2)意味着在随机缺失机制中，数据的缺失不是完全随机缺失的，缺失的概率依赖于完全观测的变量。这种类型的缺失通常对研究结果的影响较大，这种情况下，可以使用其他已有的变量来推断缺失数据。

(3) 非随机缺失

当数据缺失的概率既取决于观测到的变量，又取决于原则上应该得到的具体缺失值时，称为非随机缺失。这种缺失模式是最难处理的，因为缺失的数据无法通过其他已有的变量进行估计或预测。例如，在一项调查中，某些参与者可能选择不回答关于收入的问题，而这些缺失的数据与收入本身有关；在医疗记录中，医生可能会因为某个患者的病情特殊而经常漏填某些重要的检查结果；在金融交易中，交易员可能会因为某个客户的信用状况较差而经常遗漏某些交易信息等等。这种类型的缺失通常对研究结果的影响较大。非随机缺失的数学表达式为：

$$P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}|\boldsymbol{Y}_o, \boldsymbol{Y}_m) \tag{2.3}$$

在这种情况下，无法准确地估计或推断缺失数据。

在实践中，由于无法根据数据对 MAR、MCAR 与 NMAR 进行严格区分，本文将函数型数据的缺失模式分为点缺失(Point Missing，PM)、区间缺失(Interval Missing，IM)以及 PM/IM 混合缺失。这里 PM 和 IM 分别对应 MCAR 和 MAR(Chiou 等，2014)：

(1) 点缺失(PM)：缺失点完全独立于观测以及未观测的值，同时，缺失点被孤立、分组或随机分散。MCAR 与 MAR 均为 PM 的特殊情况。以每天每隔 5 分钟观测的交通流量数据为例，PM 缺失模式如图 2.1(左)所示。

(2) 区间缺失(IM)：区间缺失与 MAR 密切相关，但二者具有不同的重点。在函数型数据中，缺失区间指一个未被观测到的区间，而不是一个小群中未被观测到的点。同样，缺失区间通常是随机出现的。以交通流量数据为例的 IM 缺失模式如图 2.1(中)所示。

(3) PM/IM 混合缺失：PM/IM 缺失模式是 PM 和 IM 的混合。以交通流量数据为例的混合 PM/IM 缺失模式如图 2.1(右)所示。

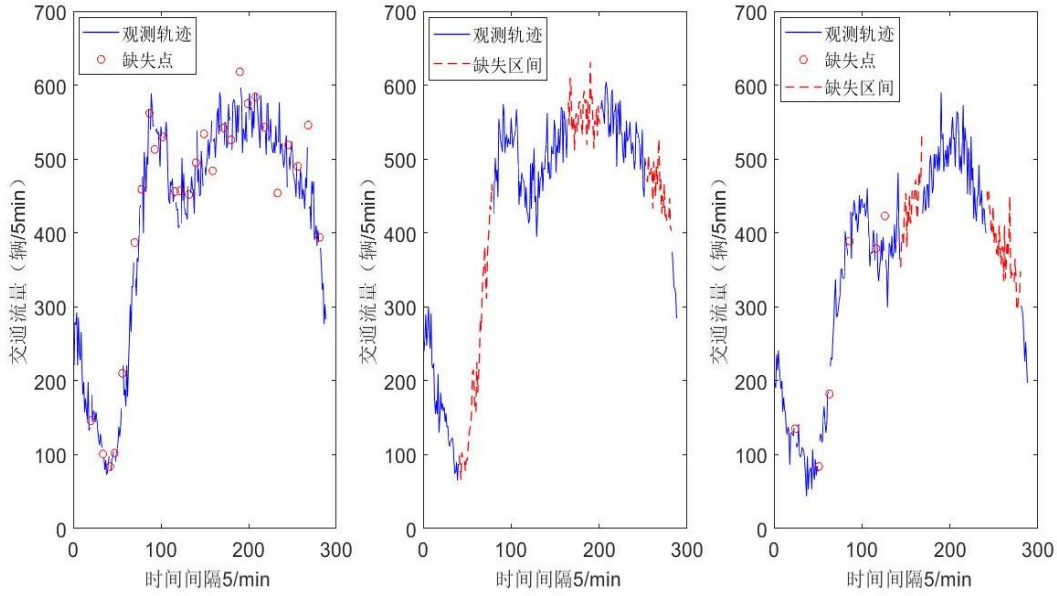


图 2.1　以交通流量数据为例展示 PM 缺失(左)、IM 缺失(中)以及混合 PM/IM 缺失(右)

2.2 函数型数据曲线拟合

函数型数据分析的关键是将采样得到的一系列离散观测值看作一个整体，利用平滑技术将其拟合成一条条连续的函数曲线，实现从有限维度(有限个点)到无限维度(无限个点的集合，具有积分、导数等更多潜在特征)的映射，能够更加系统、全面地进行统计分析。

设n维函数向量$\boldsymbol{x}(t)=[x_1(t),x_2(t),\cdots,x_n(t)]^T$为定义在连续集$\mathcal{T}$上的一个独立同分布函数型数据样本，$t\in\mathcal{T}=[a,b]$，$x_i(t)(i=1,2,\cdots,n)$是平方可积空间$\mathcal{L}^2(\mathcal{T})$上的实值曲线。假定$\tilde{y}_{ij}(i=1,2,\cdots,n,\ j=1,2,\cdots,m_i)$为函数$x_i(t)$带噪音的第$j$个离散观测，由以下一般形式的回归模型生成

$$\tilde{y}_{ij}=x_i(t_{ij})+\varepsilon_{ij} \tag{2.4}$$

$x_i(t)$可以在有限维度下近似表述为

$$x_i(t)\approx\sum_{l=1}^{r}\alpha_{il}\phi_{il}(t)=\boldsymbol{\phi}_i(t)^T\boldsymbol{\alpha}_i \tag{2.5}$$

其中，$\boldsymbol{\phi}_i(t)=[\phi_{i1}(t),\cdots,\phi_{ir}(t)]^T$是既定空间中的一组基，$\boldsymbol{\alpha}_i=[\alpha_{i1},\alpha_{i2},\cdots,\alpha_{ir}]^T$是待估系数列向量。为便于表述，将式(2.5)和式(2.4)改写为矩阵形式[①]

[①] 需要指出的是，无论每一条曲线i的观测数据点数量m_i是否相等，总可以通过数据的稀疏化排列，使得原始数据具有相同的维度m，$m\geq\max_i(m_i)$，进而可以得到式(2.4)和式(2.5)的矩阵表述。

$$\boldsymbol{x}(t) = \boldsymbol{A}^T \boldsymbol{\phi}(t) \tag{2.6}$$

$$\tilde{\boldsymbol{Y}} = \boldsymbol{\Phi}\boldsymbol{A} + \boldsymbol{E} \tag{2.7}$$

在式(2.6)中，$\boldsymbol{\phi}(t)$是一组可以统一表示各样本曲线的基函数，式(2.6)表明，曲线向量$\boldsymbol{x}(t)$中元素之间的差异由系数矩阵$\boldsymbol{A}$决定。在式(2.7)中，$\tilde{\boldsymbol{Y}}$为函数型数据矩阵，$\boldsymbol{\Phi} \in \mathbb{R}^{m\times r}$为基矩阵，$\boldsymbol{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n] \in \mathbb{R}^{r\times n}$为待估系数列向量形成的矩阵，$\boldsymbol{E}$为误差矩阵。

同理，对于多视角函数型数据，随机函数$\boldsymbol{X}(t) \in L^2(\mathcal{T})$描述样本曲线轨迹。对于有$v$个视角的样本数据，设$\boldsymbol{X}(t) = (\boldsymbol{X}_1(t), \cdots, \boldsymbol{X}_{n_v}(t))$为$L^2(\mathcal{T})$中的多视角函数型数据，其中$\boldsymbol{X}_v(t) = (x_1^v(t), \cdots, x_i^v(t), \cdots, x_n^v(t))$，$i = 1, 2, \cdots, n$且$n$为第$v$个视角的样本曲线数。对于$\boldsymbol{X}(t)$中第$v$个视角的第$i$条样本曲线$x_i^v(t)$，其第$j$个离散观测值$\tilde{y}_{ij}^v$由回归模型式(2.8)生成

$$\tilde{y}_{ij}^v = x_i^v(t_{ij}) + \epsilon_{ij}^v \tag{2.8}$$

其中ϵ_{ij}^v是具有零均值、同方差$Var\{\varepsilon_{ij}^v\} = \sigma^2$的不相关随机误差，$j = 1, 2, \cdots, m_i$为时间间隔$\mathcal{T}$上的观测时间点。进一步，利用函数型数据曲线拟合，$x_i^v(t)$可表示为

$$x_i^v(t) \approx \sum_{l=1}^{r} \alpha_{il}^v \phi_{il}^v(t) = (\boldsymbol{\phi}_i^v(t))^T \boldsymbol{\alpha}_i^v \tag{2.9}$$

其中$\boldsymbol{\phi}_i^v(t) = (\phi_{i1}^v(t), \phi_{i2}^v(t), \cdots, \phi_{ir}^v(t))^T$和$\boldsymbol{\alpha}_i^v = (\alpha_{i1}^v, \alpha_{i2}^v, \cdots, \alpha_{ir}^v)^T$分别是维数为$r$的基函数和系数向量。为便于表述，式(2.8)和式(2.9)可写为矩阵形式

$$\tilde{\boldsymbol{Y}}_v^+ = \boldsymbol{\Phi}_v \boldsymbol{A}_v + \boldsymbol{E}_v \tag{2.10}$$

$$\tilde{\boldsymbol{X}}_v^+ \approx \boldsymbol{\Phi}_v \boldsymbol{A}_v \tag{2.11}$$

其中$\boldsymbol{\Phi}_v = (\boldsymbol{\phi}_1^v(t), \cdots, \boldsymbol{\phi}_{m_v}^v(t)) \in \mathbb{R}^{m_v\times r}$，$\boldsymbol{A}_v = (\boldsymbol{\alpha}_1^v, \cdots, \boldsymbol{\alpha}_n^v)$，$\boldsymbol{E}_v = (\boldsymbol{\epsilon}_1^v, \cdots, \boldsymbol{\epsilon}_n^v)$，$\boldsymbol{\epsilon}_i^v = (\epsilon_{i1}^v, \epsilon_{i2}^v, \cdots, \epsilon_{im_v}^v)$。设$\tilde{\boldsymbol{Y}}_v \in \mathbb{R}_+^{m_v\times n}$表示第$v$个视角的原始观测数据矩阵，$n$、$m_v$分别为第$v$个视角观测数据的样本数以及观测维数。

2.3 基于非负矩阵分解的矩阵填充

对于大规模矩阵恢复问题，一类经典的矩阵填充方法是基于矩阵分解，其中低秩矩阵通过系数矩阵和基矩阵的乘积来近似。矩阵分解的基本目标函数如下：

$$\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{O} = ||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T|| + \mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) \tag{2.12}$$

其中，$\boldsymbol{X} \in \mathbb{R}^{m\times n}$是要近似的数据矩阵，$\boldsymbol{U} \in \mathbb{R}^{m\times d}$、$\boldsymbol{V} \in \mathbb{R}^{n\times d}$是两个低维矩阵$(d < \min(m,n))$，$\mathcal{L}(\boldsymbol{U},\boldsymbol{V})$为避免过拟合的正则化项。然而在实际问题中，诸如图像数据、文本数据等，具有非负性要求。为了适应于非负数据的处理，Lee & Seung(1999)将非负性约束引入矩阵分解中，提出了 NMF 算法，即将非负原始数据矩阵$\boldsymbol{X}$分解成两个非负低秩矩阵$\boldsymbol{U}$、$\boldsymbol{V}$的乘积，得到如下优化问题

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{O} = ||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T||_F^2 + \mathcal{L}(\boldsymbol{U},\boldsymbol{V}) \\ &\text{s.t.} \qquad \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0 \end{aligned} \tag{2.13}$$

目标函数式(2.13)同时针对待估参数矩阵$\boldsymbol{U}$、$\boldsymbol{V}$时是非凸函数，难以获得全局最优解，根据 Lee & Seung(1999)提出的迭代更新算法可找到局部最小值。$\boldsymbol{U}$、$\boldsymbol{V}$的更新规则分别为

$$\begin{aligned} \boldsymbol{U}_{ij}^{t+1} &\leftarrow \boldsymbol{U}_{ij}^{t} \frac{(\boldsymbol{X}\boldsymbol{V})_{ij}}{(\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{V})_{ij}} \\ \boldsymbol{V}_{ij}^{t+1} &\leftarrow \boldsymbol{V}_{ij}^{t} \frac{(\boldsymbol{X}^T\boldsymbol{U})_{ij}}{(\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{U})_{ij}} \end{aligned} \tag{2.14}$$

当矩阵$\boldsymbol{X}$包含大规模缺失时，式(2.13)修正为

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{O} = ||\boldsymbol{O} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)||_F^2 + \mathcal{L}(\boldsymbol{U},\boldsymbol{V}) \\ &\text{s.t.} \qquad \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0 \end{aligned} \tag{2.15}$$

其中，$\odot$是矩阵的哈达玛积，$\boldsymbol{O} \in \mathbb{R}^{m\times n}$是与$\boldsymbol{X}$同型的投影矩阵，即若$\boldsymbol{X}$中的条目$\boldsymbol{X}_{ij}$可观测，则$\boldsymbol{O}_{ij} = 1$；否则$\boldsymbol{O}_{ij} = 0$。求解式(2.15)得到$\boldsymbol{U}$、$\boldsymbol{V}$的更新规则，从而得到$\boldsymbol{X}$的估计，完成矩阵填充过程。

2.4 最优图正则化项

通过使用非负约束，NMF 可以学习数据中基于部分的表示。然而，NMF 在欧氏空间中执行这种学习，未能发现数据空间的几何结构，为了充分利用这一信息，引入图正则化的概念。

(1) 一阶拉普拉斯矩阵

NMF 试图找到一个可以对数据的线性近似进行优化的基，其中数据的分布为P_X。根据流行假设，设$f_k(\boldsymbol{x}_i) = v_{ik}$是生成原始数据点$\boldsymbol{x}_i$到轴$\boldsymbol{u}_k$映射的函数，利用$||f_k||_M^2$测量$f_k$在数据固有几何结构中沿测地线的平滑度。考虑数据是紧致子流形$\mathcal{M} \subset \mathbb{R}^m$的情况时，有

$$||f_k||_M^2 = \int_{\boldsymbol{x}\in\mathcal{M}} ||\nabla_{\mathcal{M}} f_k||^2 dP_X(\boldsymbol{x}) \tag{2.16}$$

其中，$\nabla_{\mathcal{M}}$是f_k沿着流行$\mathcal{M}$的梯度。事实上，数据流行通常是未知的，因此，式(2.16)中的$||f_k||_M^2$无法计算，而对谱图理论和流形学习理论的研究表明：$||f_k||_M^2$可以通过数据点分散上的最近邻图离散地近似。

考虑一个具有n个顶点的图，其中每个顶点对应于一个数据点。定义边权重矩阵（相似矩阵）

$$\boldsymbol{W}_{1_{ij}} = \begin{cases} 1, & \text{if } \ \boldsymbol{x}_i \in N_k(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_k(\boldsymbol{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{2.17}$$

其中，$N_k(\boldsymbol{x}_i)$表示$\boldsymbol{x}_i$的k个最近邻点，式(2.17)表示若$\boldsymbol{x}_i \in N_k(\boldsymbol{x}_j)$，则矩阵$\boldsymbol{W}_1$中元素$\boldsymbol{W}_{1_{ij}} = 1$，反之为 0。定义$\boldsymbol{L}_1 = \boldsymbol{D}_1 - \boldsymbol{W}_1$，其中$\boldsymbol{D}_1$为对角矩阵，每一个对角元素为$\boldsymbol{W}_1$的列和，即第$i$个对角线元素$\boldsymbol{D}_{1_{ii}} = \sum_{j=1}^{n} \boldsymbol{W}_{1_{ij}}$，$\boldsymbol{L}_1$被称为图拉普拉斯算子，该矩阵是$\nabla_{\mathcal{M}}$在流行上的离散近似，因此，$||f_k||_M^2$的离散近似可以计算为

$$\begin{aligned} \mathcal{R}_k &= \frac{1}{2}\sum_{i,j=1}^{n} (f_k(\boldsymbol{x}_i) - f_k(\boldsymbol{x}_j))^2 \boldsymbol{W}_{1_{ij}} \\ &= \sum_{i=1}^{n} f_k(\boldsymbol{x}_i)\boldsymbol{D}_{1_{ii}} - \sum_{i,j=1}^{n} f_k(\boldsymbol{x}_i) f_k(\boldsymbol{x}_j)\boldsymbol{W}_{1_{ij}} \\ &= \sum_{i=1}^{n} v_{ik}^2 \boldsymbol{D}_{1_{ii}} - \sum_{i,j=1}^{n} v_{ik} v_{jk} \boldsymbol{W}_{1_{ij}} \\ &= \boldsymbol{v}_k^T \boldsymbol{D}_1 \boldsymbol{v}_k - \boldsymbol{v}_k^T \boldsymbol{W}_1 \boldsymbol{v}_k \\ &= \boldsymbol{v}_k^T \boldsymbol{L}_1 \boldsymbol{v}_k \end{aligned} \tag{2.18}$$

则$\mathcal{R}_k$可以用于测量映射函数f_k沿着数据集的内在几何中的测地线的平滑度。有

$$\mathcal{J}_{\text{manifold}} = \sum \mathcal{R}_k = \text{tr}\left(\boldsymbol{V}^T \boldsymbol{L}_1 \boldsymbol{V}\right) \tag{2.19}$$

传统矩阵填充方法以及 MVNFMC 方法在估计缺失数据时，虽然考虑了输入数据的整体结构信息，但忽视了输入数据的内部几何结构信息。Cai 等(2008)的研究表明，拉普拉斯图可以用于发现原始数据中的几何结构，且保持样本点在原始高维特征空间中的邻近关系不变。因此，本文在处理缺失数据时，引入图正则化，进一步提高插补方法的有效性。

(2) 二阶拉普拉斯矩阵

近年来，随着图卷积神经网络(GCNN)(Defferrard 等，2016)的普及，高阶相

似信息受到越来越多的关注。众所周知，一阶和二阶连接是图论中的基本概念(Tang 等，2015)。具体来说，在图嵌入中，两点之间的相似度通过一阶关系表示。但如果两个顶点具有相同的邻域结构，即两个顶点具有共同的一阶连接点，则这两个顶点具有二阶连接，二阶相似度的定义如下。

定义 2.1(二阶相似性) 网络中两个顶点之间的二阶相似性是它们的邻域网络结构之间的相似性(Tang 等，2015)。

根据二阶相似性，$\boldsymbol{w}_j$是一阶相似矩阵$\boldsymbol{W}_1$的第j列，对应的二阶相似矩阵$\boldsymbol{W}_2$为

$$W_{2_{ij}} = \boldsymbol{w}_i^T \boldsymbol{w}_j$$

则二阶拉普拉斯矩阵$\boldsymbol{L}_2 = \boldsymbol{D}_2 - \boldsymbol{W}_2$，度矩阵$\boldsymbol{D}_2$的对角元素为$D_{2_{ii}} = \sum_{j=1}^{n} W_{2_{ij}}$。研究表明，在实际问题中，仅考虑一阶相似性不足以挖掘缺失数据的几何结构(Tang 等，2015)。然而，大多数现有插补方法都没有考虑到高阶信息的作用。因此，本文引入最优图正则化，同时考虑一阶、二阶信息，充分捕捉数据的邻域几何结构，提高缺失数据的插补精度。定义最优拉普拉斯矩阵

$$\boldsymbol{L} = \alpha_1 \boldsymbol{L}_1 + \alpha_2 \boldsymbol{L}_2 \tag{2.20}$$

其中，α_1, α_2为平衡参数，则最优图正则化可以表示为

$$\min_{\boldsymbol{V} \geq 0} \operatorname{tr}(\boldsymbol{V}^T \boldsymbol{L} \boldsymbol{V}) \tag{2.21}$$

2.5 希尔伯特-施密特独立性准则

希尔伯特-施密特独立性准则(Hilbert-Schmidt Independence Criterion，HSIC)最早由 Gretton 等(2005)提出，主要用来衡量任意两个空间上两个变量之间的依赖性关系。给定两个可分离的再生核希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)$\mathcal{F}$、$\mathcal{G}$，对于随机变量x、y，有映射函数$\phi : x \to \mathcal{F}$和$\varphi : y \to \mathcal{G}$。定义 HSIC 为交叉协变量$C_{xy}$上的希尔伯特范数，有

$$\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) = ||C_{xy}||_{HS}^2 = ||E\left((\phi(x) - \mu(x)) \otimes (\varphi(y) - \mu(y))\right)||_{HS}^2 \tag{2.22}$$

其中，$\mu(x) = E(\phi(x))$和$\mu(y) = E(\varphi(y))$分别为x、y的期望，p_{xy}为x和y的联合测度。

Gretton 等(2005)给出了 HSIC 的估计量。设$\boldsymbol{Z} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \cdots, (\boldsymbol{X}_N, \boldsymbol{Y}_N)\}$来自联合测度$p_{xy}$，则$\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N)$与$\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N)$之间的经验 HSIC 定义如下：

$$\mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{(N-1)^2}\mathrm{tr}(\boldsymbol{HK_XHK_Y}) \tag{2.23}$$

其中，$\boldsymbol{H}=\boldsymbol{I}-(1/N)\boldsymbol{N}$，$\boldsymbol{I}$表示单位矩阵，$\boldsymbol{N}\in\mathbb{R}^{N\times N}$的元素均为 1，$\boldsymbol{K_X}$为核矩阵。

为便于理解 HSIC，简要介绍经验 HSIC 的矩阵运算形式。根据定义(2.23)可知

$$\boldsymbol{H}=\begin{pmatrix}1-\frac{1}{N} & -\frac{1}{N} & \cdots & -\frac{1}{N}\\ -\frac{1}{N} & 1-\frac{1}{N} & \cdots & -\frac{1}{N}\\ \vdots & \vdots & \ddots & -\frac{1}{N}\\ -\frac{1}{N} & -\frac{1}{N} & \cdots & 1-\frac{1}{N}\end{pmatrix} \tag{2.24}$$

$$\boldsymbol{K_X}=\begin{pmatrix}K_{11} & K_{12} & \cdots & K_{1N}\\ K_{21} & K_{22} & \cdots & K_{2N}\\ \vdots & \vdots & \ddots & \vdots\\ K_{N1} & K_{N2} & \cdots & K_{NN}\end{pmatrix} \tag{2.25}$$

有

$$\boldsymbol{K_XH}=\begin{pmatrix}K_{11}-\overline{K1} & K_{12}-\overline{K}_1 & \cdots & K_{1N}-\overline{K}_1\\ K_{21}-\overline{K}_2 & K_{22}-\overline{K}_2 & \cdots & K_{2N}-\overline{K}_2\\ \vdots & \vdots & \ddots & \vdots\\ K_{N1}-\overline{K}_N & K_{N2}-\overline{K}_N & \cdots & K_{NN}-\overline{K}_N\end{pmatrix} \tag{2.26}$$

根据定义(2.23)可以看出，相比其他的相关性估计方法，HSIC 的经验估计计算简单，且 HSIC 可以测量任意两个空间中变量之间的相关性，既包括线性相关、也包括非线性。这是 HSIC 相比其他方法最显著的优势。此外，经验估计值以$1/\sqrt{N}$的速率收敛到总体估计，N为样本量，因此基于 HSIC 的独立性测试不存在学习速率慢的问题，特别地，随着样本量的增加，可以保证以较高的概率检测到任何现有的相关性。

2.6 函数型矩阵填充的一般框架

函数型矩阵填充问题的一般框架以曲线拟合为基础，旨在求解式(2.7)中系数矩阵$\boldsymbol{A}$的估计，可以采用最小化目标函数

$$\arg\min_{\boldsymbol{A}}||\boldsymbol{O}\odot(\tilde{Y}-\boldsymbol{\Phi A})||_F^2+\lambda\mathrm{Pen}(\boldsymbol{A}) \tag{2.27}$$

得到结果，进而由式(2.6)可得曲线$\boldsymbol{x}(t)$的估计。其中，$\mathrm{Pen}(\boldsymbol{A})$为惩罚项，$\lambda$为旋钮参数。根据这一思想，Kidziński & Hastie(2018)提出软函数型数据填充法(SFI)，

目标函数为

$$\arg\min_{\boldsymbol{A}} \left\| \boldsymbol{O} \odot \left(\tilde{\boldsymbol{Y}} - \boldsymbol{\Phi}\boldsymbol{A} \right) \right\|_F^2 + \lambda \|\boldsymbol{A}\|_0 \tag{2.28}$$

针对式(2.28)进行奇异值分解迭代，从稀疏观测值中求解参数矩阵$\boldsymbol{A}$、预测曲线轨迹，进而开展数据修复工作。进一步地，Kidziński & Hastie(2018)将惩罚项由核范数替换为l_0范数，提出了硬函数型数据填充法(HFI)，目标函数如下

$$\arg\min_{\boldsymbol{A}} \left\| \boldsymbol{O} \odot \left(\tilde{\boldsymbol{Y}} - \boldsymbol{\Phi}\boldsymbol{A} \right) \right\|_F^2 + \lambda \|\boldsymbol{A}\|_0 \tag{2.29}$$

求解式(2.28)和式(2.29)，为函数型矩阵填充提供了两种实现方法。然而，当面临非负函数型数据时，以上插补方法不能保证结果非负。在 Kidziński & Hastie(2018)方法的基础上，本文的后续讨论以非负函数型数据为研究对象，构造函数型数据填充方法，用以解决特定情形下的函数型数据填充问题。

2.7 缺失插补评价指标

在缺失处理问题中，插补方法的性能评估可利用不同的评价指标。对于样本曲线$\boldsymbol{x}_i(t)$和估计曲线$\hat{\boldsymbol{x}}_i(t)(i = 1, 2, \cdots, n)$，本文主要使用的评价指标有：均方根误差(Root Mean Square Error，RMSE)、平均绝对误差(Mean Absolute Error ，MAE)、平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)、归一化均方根误差(Normalized Root Mean Square Error，NRMSE)以及相关系数。

(1) 均方根误差

计算插补值与缺失值之间误差平方和的均值，用于衡量插补值与缺失值的匹配程度，其值越小说明插补效果越好，定义

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t) \right)^2} \tag{2.30}$$

(2) 平均绝对误差

表示插补值与缺失值之间绝对误差的平均值，用于评估预测结果和真实数据集的接近程度，其值越小说明插补效果越好，定义

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| \boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t) \right| \tag{2.31}$$

(3) 平均绝对百分比误差

MAPE 的值越小，说明模型插补效果越好，具有更高的插补精度，定义

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)}{\boldsymbol{x}_i(t)}\right| \tag{2.32}$$

(4) 归一化均方根误差

在均方根误差的基础上进行归一化处理(除以均值函数)，使 NRMSE 在$(0,1)$之间取值。定义

$$\mathrm{NRMSE}=\frac{1}{\overline{\boldsymbol{x}}_i(t)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\right)^2} \tag{2.33}$$

(5) 相关系数

用来衡量变量之间的联系或依赖程度，取值范围为$[-1,1]$。插补值与真实值的相关系数越大，则插补值的有效性越高。相关系数的计算公式为

$$r(\boldsymbol{x}_i(t),\hat{\boldsymbol{x}}_i(t))=\frac{Cov(\boldsymbol{x}_i(t),\hat{\boldsymbol{x}}_i(t))}{\sigma_{\boldsymbol{x}_i(t)}\sigma_{\hat{\boldsymbol{x}}_i(t)}} \tag{2.34}$$

其中，$Cov(x_i(t),\hat{x}_i(t))=E\left((x_i(t)-\overline{x}_i(t))(\hat{x}_i(t)-\overline{\hat{x}}_i(t))\right)$。

3 融合类信息的函数型矩阵填充方法

3.1 方法框架

为了处理非负函数型数据缺失插补问题，本文在函数型数据分析的框架下，充分考虑不同样本之间的潜在差异、挖掘样本之间的相关性，提出融合类信息的函数型矩阵填充方法(CNFMC)。首先构造函数型矩阵填充模型进行初始插补，获取完整数据集；其次利用聚类分析探索相关性强的同质子群，将具有相似变化轨迹的样本划分到同一类中；然后对每一类样本利用类间信息进行插补；最后利用集成学习方法将不同类别下每条样本的插补结果进行融合，得到最终的插补值。CNFMC 方法的具体实现步骤如下：

步骤 1：构建函数型矩阵填充方法(Non-negative Functional Matrix Completion，NFMC)对含缺失的观测矩阵$\boldsymbol{Y}$进行初始插补。

针对非负函数型数据缺失情况，引入非负约束，融合非负矩阵分解、矩阵填充等思想，构建函数型矩阵填充方法 NFMC。NFMC 方法首先通过曲线拟合，将无穷维问题转化为有限维问题；其次对系数矩阵进行非负矩阵分解；并采用函数型数据分析框架进行数据补齐。需要指出的是，以上 3 个策略不是串联进行的，而是在一个统一的目标函数框架中进行求解。建立目标函数

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0}\left\{\frac{1}{2}\|\boldsymbol{O}\odot(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{T})\|_F^2\right\} \tag{3.1}$$

在此基础上，为了求解的稳定，对$\boldsymbol{U}$和$\boldsymbol{V}$施加 Frobenius 范数惩罚项，建立 NFMC 方法的目标函数

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0}\mathcal{O}=\left\{\frac{1}{2}\|\boldsymbol{O}\odot(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{T})\|_F^2+\frac{\alpha}{2}\|\boldsymbol{U}\|_F^2+\frac{\beta}{2}\|\boldsymbol{V}\|_F^2\right\} \tag{3.2}$$

其中，$\boldsymbol{Y}\in\mathbb{R}^{m\times n}$为函数型数据的离散观测矩阵，$\boldsymbol{O}\in\mathbb{R}^{m\times n}$是与$\boldsymbol{Y}$同型的投影矩阵，即若$\boldsymbol{Y}$中的条目可观测，则$\boldsymbol{O}_{ij}=1$；否则$\boldsymbol{O}_{ij}=0$。$\boldsymbol{\Phi}\in\mathbb{R}^{m\times r}$为通过 B-样条基函数形成的基矩阵，$\boldsymbol{U}\in\mathbb{R}_{+}^{r\times d}$为 NMF 的非负基矩阵，$\boldsymbol{V}\in\mathbb{R}_{+}^{n\times d}$为 NMF 的非负系数矩阵，$m$为原始数据的变量数(维数)，$n$为样本量，$r$为曲线拟合基函数的数量，$d$为 NMF 的秩。

步骤 2：通过聚类划分引入样本类信息，使得每一类样本相似度高，并借助类内样本相关性插补缺失值。

借助聚类结果挖掘函数型数据样本曲线的相关性，进一步提升插补精度。由于每一类中的样本相关性较强、具有相似的变化轨迹，故在每一簇样本中利用NFMC 方法进行局部缺失填充，类内样本的高相关性有助于提高 NFMC 方法的插补精度。具体地，①对步骤 1 中的矩阵$\boldsymbol{V}$应用函数型聚类算法(FNMF)(高海燕等， 2020)，将样本划分为k类($k=1,2,\cdots,K$)，设最终聚类结果为$\{C_1,C_2,\cdots,C_K\}$；②对每个同质子群C_i应用 NFMC 方法，得到对应第i类的插补结果$\hat{\boldsymbol{Y}}_k^i$ $(1\leq i\leq k)$，并将k个插补结果$\{\hat{\boldsymbol{Y}}_k^1,\hat{\boldsymbol{Y}}_k^2,\cdots,\hat{\boldsymbol{Y}}_k^i,\cdots,\hat{\boldsymbol{Y}}_k^k\}$排列组合为$\hat{\boldsymbol{Y}}_k$；③重复过程①和②，直到$k=K$。最终得到$\boldsymbol{Y}$的多个插补结果，记为$\{\hat{\boldsymbol{Y}}_1,\hat{\boldsymbol{Y}}_2,\cdots\hat{\boldsymbol{Y}}_i,\cdots,\hat{\boldsymbol{Y}}_K\}$。

步骤 3：采用自加权集成学习算法动态赋权重计算出最终插补值。

不同的聚类数k会得到不同的插补结果，为充分利用不同聚类里的插补结果以及降低k的影响，本文采用自加权集成学习算法动态赋权重对K个插补结果$\{\hat{\boldsymbol{Y}}_1,\hat{\boldsymbol{Y}}_2,\cdots\hat{\boldsymbol{Y}}_i,\cdots,\hat{\boldsymbol{Y}}_K\}$进行融合，得到最终的插补结果$\hat{\boldsymbol{Y}}$。以第$l$条样本曲线为例，将样本$\boldsymbol{Y}^{\cdot l}$的$K$个插补结果$\{\hat{\boldsymbol{Y}}_1^{\cdot l},\hat{\boldsymbol{Y}}_2^{\cdot l},\cdots\hat{\boldsymbol{Y}}_i^{\cdot l},\cdots,\hat{\boldsymbol{Y}}_K^{\cdot l}\}$进行融合得到最终的插补结果$\hat{\boldsymbol{Y}}^{\cdot l}$，满足

$$\underset{\hat{\boldsymbol{Y}}^{\cdot l}}{\operatorname{argmin}}\sum_{k=1}^{K}\omega_k||\boldsymbol{Y}^{\cdot l}-\hat{\boldsymbol{Y}}_k^{\cdot l}||_2 \tag{3.3}$$

其中，ω_k为权重，定义$\omega_k=1/2(\frac{1}{N}||\boldsymbol{Y}-\hat{\boldsymbol{Y}}_k||_F^2)$，根据聚类数为$k$时插补值与真实值的误差动态调整。求解式(3.3)的优化问题，得样本$\boldsymbol{Y}^{\cdot l}$的最终插补结果为

$$\hat{\boldsymbol{Y}}^{\cdot l}=\frac{\sum_{k=1}^{K}\omega_k\hat{\boldsymbol{Y}}_k^{\cdot l}}{\sum_{k=1}^{K}\omega_k} \tag{3.4}$$

CNFMC 方法在 NFMC 方法初始插补的基础上，考虑样本曲线的相关性，嵌入类内信息，使得类内样本相关性强，在有效利用全局数据特征的同时保留更多的局部信息。与 NFMC 方法利用全局数据特征信息插补缺失值的优异表现相结合，CNFMC 方法由于 NMF 的非负约束性与类内局部特征信息的提取，相比其他插补方法有更好的解释性和准确性。此外，CNFMC 方法采用自加权集成学习算法，不仅降低了聚类数k对插补结果的影响，还对每条样本的多个插补结果进行融合，有助于进一步提高插补的有效性。

3.2 求解算法

3.2.1 求解过程

目标函数式(3.2)同时针对待估参数矩阵$\boldsymbol{U}$和$\boldsymbol{V}$时是非凸函数，难以获得全局最优解。为此，采用乘性迭代方法(Liang 等，2020)，并利用 KKT 互补松弛条件来施加非负性约束，提出一种获得局部最优解的交替迭代算法。

令目标函数式(3.2)的拉格朗日函数为

$$\mathcal{L}=\frac{1}{2}\|\boldsymbol{O}\odot(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\|_F^2+\frac{\alpha}{2}\|\boldsymbol{U}\|_F^2+\frac{\beta}{2}\|\boldsymbol{V}\|_F^2-\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{U}^T)-\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{V}^T) \tag{3.5}$$

在此基础上，依次更新求解$\boldsymbol{U}$和$\boldsymbol{V}$，具体求解的更新规则如下。

(1) 保持$\boldsymbol{V}$不变，更新$\boldsymbol{U}$。

对式(3.5)关于$\boldsymbol{U}$求偏导，并令$\frac{\partial\mathcal{L}}{\partial\boldsymbol{U}}=0$，结合哈达玛积和迹的性质(张贤达，2013)，即

$$(\boldsymbol{A}\odot\boldsymbol{B})^T=\boldsymbol{A}^T\odot\boldsymbol{B}^T,\ \boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{A}=\boldsymbol{O}\odot\boldsymbol{A}$$

$$\mathrm{tr}[(\boldsymbol{O}^T\odot\boldsymbol{A}^T)(\boldsymbol{O}\odot\boldsymbol{A})]=\mathrm{tr}[\boldsymbol{A}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{A})]$$

则有

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{U}}&=\frac{\frac{1}{2}\partial\ \mathrm{tr}\left(\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-2\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}\right)}{\partial\boldsymbol{U}}\\&\quad+\frac{\partial\ \mathrm{tr}\left(\frac{\alpha}{2}\boldsymbol{U}^T\boldsymbol{U}-\boldsymbol{\Lambda}\boldsymbol{U}^T\right)}{\partial\boldsymbol{U}}\\&=\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}+\alpha\boldsymbol{U}-\boldsymbol{\Lambda}\\&=0\end{aligned}$$

从而可得

$$\boldsymbol{\Lambda}=\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}+\alpha\boldsymbol{U}$$

非负约束使得 KKT 条件$\Lambda\odot\boldsymbol{U}=0$成立，即满足

$$[\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}+\alpha\boldsymbol{U}]_{ij}\odot\boldsymbol{U}_{ij}=0 \tag{3.6}$$

式(3.6)对j求和，依据矩阵乘法，有

$$\begin{aligned}&\sum_j[\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}+\alpha\boldsymbol{U}]_{ij}\odot\boldsymbol{U}_{ij}\\&=(\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}-\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V})_{ii}\\&=0\end{aligned}$$

可得$\boldsymbol{U}$的更新规则

$$\boldsymbol{U}_{ij} \leftarrow \boldsymbol{U}_{ij}\sqrt{\frac{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V})_{ij}}{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}+\alpha\boldsymbol{U})_{ij}}} \tag{3.7}$$

(2) 保持$\boldsymbol{V}$不变，更新$\boldsymbol{U}$。

式(3.5)关于$\boldsymbol{V}$求偏导，并令$\frac{\partial\mathcal{L}}{\partial\boldsymbol{V}}=0$，有

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{U}} &= \frac{\frac{1}{2}\partial\ \mathrm{tr}\left(\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)-2\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T(\boldsymbol{O}^T\odot\boldsymbol{O}^T\odot\boldsymbol{Y}^T)\right)}{\partial\boldsymbol{U}} \\ &\quad + \frac{\partial\ \mathrm{tr}\left(\frac{\beta}{2}\boldsymbol{V}^T\boldsymbol{V}-\boldsymbol{\Gamma}\boldsymbol{V}^T\right)}{\partial\boldsymbol{U}} \\ &= (\boldsymbol{O}\odot\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{\Phi}^T)\boldsymbol{\Phi}\boldsymbol{U}-(\boldsymbol{O}^T\odot\boldsymbol{Y}^T)\boldsymbol{\Phi}\boldsymbol{U}+\beta\boldsymbol{V}-\boldsymbol{\Gamma} \\ &= 0\end{aligned}$$

同理，可得$\boldsymbol{V}$的更新规则为

$$\boldsymbol{V}_{ij} \leftarrow \boldsymbol{V}_{ij}\sqrt{\frac{((\boldsymbol{O}^T\odot\boldsymbol{Y}^T)\boldsymbol{\Phi}\boldsymbol{U})_{ij}}{((\boldsymbol{O}^T\odot\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{\Phi}^T)\boldsymbol{\Phi}\boldsymbol{U}+\beta\boldsymbol{V})_{ij}}} \tag{3.8}$$

综上分析，依次根据式(3.7)和式(3.8)交替迭代更新$\boldsymbol{U}$和$\boldsymbol{V}$，可以完成优化问题式(3.2)的求解，即实现了 NFMC 方法，进而通过步骤 2、3，完成具体数据填充工作。NFMC 方法的具体执行过程如算法 3.1 所示。

算法 3.1 NFMC 方法

输入：非负函数型数据矩阵$\boldsymbol{Y}$，基矩阵$\boldsymbol{\Phi}$，α，β，p，$\varepsilon_0=10^{-7}$，$\varepsilon_1=10^{-4}$，最大更新迭代次数$=1000$。

过程：

1：初始化：随机生成基矩阵$\boldsymbol{U}_0$和系数矩阵$\boldsymbol{V}_0$；

2：**for** $t=1,2,\cdots$最大更新迭代次数

3：固定$\boldsymbol{V}$，根据式(3.7)更新$\boldsymbol{U}_{t+1}$；

4：固定$\boldsymbol{U}$，根据式(3.8)更新$\boldsymbol{V}_{t+1}$；

5： **if** $\|\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^T-\boldsymbol{U}_t\boldsymbol{V}_t^T\|_F^2/\|\boldsymbol{U}_t\boldsymbol{V}_t^T\|_F^2\le\varepsilon_0$ **and** $|\mathcal{O}_{t+1}-\mathcal{O}_t|/\mathcal{O}_t\le\epsilon_1$

6： **break**

7： **end if**

8：**end for**

输出：基矩阵$\boldsymbol{U}_{t+1}$和系数矩阵$\boldsymbol{V}_{t+1}$。

3.2.2 收敛性证明

算法 3.1 是局部收敛的，下面讨论 NFMC 方法的收敛性。

定理 3.1 目标函数式(3.2)在更新规则式(3.7)和(3.8)下是单调递减的。

证明 采用标准辅助函数法(Lee & Seung，2001)证明更新规则式(3.7)的收敛性。在目标函数式(3.2)中，剔除无关项，保留与$\boldsymbol{U}$有关的项，有

$$\mathcal{L}(\boldsymbol{U}) := \mathrm{tr}\Big(\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} - 2\boldsymbol{U}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}\Big)+\alpha\mathrm{tr}(\boldsymbol{U}^T\boldsymbol{U})$$

构造$\mathcal{L}(\boldsymbol{U})$的辅助函数$\boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}_t)$(Ding 等，2006)

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}^t) = &\sum_{i,j}\frac{\big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}^t\boldsymbol{V}^T)\boldsymbol{V}\big)(i,j)\boldsymbol{U}(i,j)^2}{\boldsymbol{U}^t(i,j)} + \alpha\sum_{i,j}\frac{\boldsymbol{U}^t(i,j)\boldsymbol{U}(i,j)^2}{\boldsymbol{U}^t(i,j)}\\ &-2\sum_{i,j}\big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}\big)_{(i,j)}\boldsymbol{U}^t(i,j)\big(1+log\frac{\boldsymbol{U}(i,j)}{\boldsymbol{U}^t(i,j)}\big)\end{aligned} \tag{3.9}$$

则满足条件

$$\boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}) = \boldsymbol{L}(\boldsymbol{U}),\quad \boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}_t)\geq\boldsymbol{L}(\boldsymbol{U})$$

如果取$\boldsymbol{U}_{t+1}$使得

$$\boldsymbol{U}_{t+1} = \arg\min_t \boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}_t) \tag{3.10}$$

成立，易知$\mathcal{L}(\boldsymbol{U})$是单调递减的，即

$$\boldsymbol{L}(\boldsymbol{U}_{t+1})\leq\boldsymbol{G}(\boldsymbol{U}_{t+1},\boldsymbol{U}_t)\leq\boldsymbol{G}(\boldsymbol{U}_t,\boldsymbol{U}_t)\leq\boldsymbol{L}(\boldsymbol{U}_t)$$

根据式(3.10)，使辅助函数式(3.9)达到最小，求解$\boldsymbol{U}_{t+1}$。对式(3.9)关于$\boldsymbol{U}_{ij}$求偏导，得

$$\begin{aligned}\frac{\partial\boldsymbol{G}(\boldsymbol{U},\boldsymbol{U}_t)}{\partial\boldsymbol{U}_{ij}} = &2\frac{\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_t\boldsymbol{V}^T)\boldsymbol{V}\Big)_{ij}\boldsymbol{U}_{ij}}{\boldsymbol{U}_{t,ij}} + 2\alpha\boldsymbol{U}_{ij}\\ &-2\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}\Big)_{ij}\frac{\boldsymbol{U}_{t,ij}}{\boldsymbol{U}_{ij}}\end{aligned}$$

令$\frac{\partial G(\boldsymbol{U},\boldsymbol{U}_t)}{\partial\boldsymbol{U}_{ij}}=0$，则有

$$\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V}\Big)_{ij}\frac{\boldsymbol{U}_{t,ij}}{\boldsymbol{U}_{t+1,ij}} = \Big(\frac{\big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_t\boldsymbol{V}^T)\boldsymbol{V}\big)_{ij}}{\boldsymbol{U}_{t,ij}}+\alpha\Big)\boldsymbol{U}_{t+1,ij}$$

进一步得到

$$\boldsymbol{U}_{t+1,ij} = \boldsymbol{U}_{t,ij}\sqrt{\frac{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{Y})\boldsymbol{V})_{ij}}{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_t\boldsymbol{V}^T)\boldsymbol{V}+\alpha\boldsymbol{U}_t)_{ij}}}$$

上式即为$\boldsymbol{U}$的更新规则式(3.7)。类似地，可证得$\boldsymbol{V}$的更新规则式(3.8)。

3.2.3 计算复杂度分析

NFMC 方法的时间复杂度主要体现在$\boldsymbol{U}$和$\boldsymbol{V}$的更新迭代中。对于的$\boldsymbol{U}$的更新，需要$O(2mnr+2nrd+mrd+mnd+rd)$加法运算、$O(2mn+2mnr+2nrd+mrd+mnd+rd)$乘法运算、$O(rd)$除法运算。则在一次迭代中，更新$\boldsymbol{U}$的时间复杂度为$O(2mn+4mnr+4nrd+2mrd+2mnd+3rd)$。同理，更新$\boldsymbol{V}$的时间复杂度为$O(mn+6mnr+6nrd+3nd)$。故 NFMC 方法迭代一次的时间复杂度为$O(mn+mnr)$，当迭代次数为$t$时，时间复杂度为$O(t(mn+mnr))$。在 CNFMC 方法中，当执行$K$个 NFMC 局部填充时，时间复杂度为$O(Kt(mn+mnr))$。

3.3 模拟实验

3.3.1 数据集

为了评估 CNFMC 方法的插补性能，本文在实例数据集中进行模拟插补实验。数据来源于公共交通数据集 PeMS(http://pems.dot.ca.gov.)美国加州路网中 6 个检测站[①]2014 年 5~6 月的交通流量数据。检测器采集数据的时间间隔$\triangle t$为 5 分钟，每天每条数据包含$m=288$个观测值。考虑到交通流量在节假日与正常工作日之间有显著差异，故不考虑节假日，只分析正常工作日的交通流量数据，为便于比较，选取不含缺失值的连续 30 天观测，交通流量数据样本总数$n=180$。

3.3.2 实验设置

实验主要通过 R 4.1.3 实现，实验的计算机环境为：Intel(R) Core(TM) i5-5200U CPU2.20 GHz，内存 6GB，Windows10 64 位操作系统。首先，人为生成带有点缺失(PM)和区间缺失(IM)的条目。具体工作如下：

(1) PM：以逐点方式对缺失的条目进行统一随机采样。

(2) IM：生成不同大小的缺失区间，这些区间在每个曲线轨迹中都有均匀分布的起始位置。缺失区间的长度遵循均匀分布。

[①] 美国加州路网中的 6 个检测站 id：716421、716424、716440、716442、718155、716453。

本文采用混合 PM/IM 模式，其中混合 PM/IM 模式将 PM 和 IM 结合用于单独的曲线轨迹；其次为每种模式生成nmp缺失点，其中n是曲线轨迹数，m是时间点数，p是缺失率并分别设定为 15%、20%、30%、40%、50%、60%和 70%。

其次，在不同缺失比例设定下，将 CNFMC 方法与下述 10 种典型插补方法进行对比，并利用 RMSE、MAE 以及 MAPE 评估每种方法的插补性能。

均值填充：基于样本的均值对缺失值进行填充。

线性插值：基于样本观测值的时间相关性对缺失值进行估计。

KNN：基于样本观测值之间的相似度估计缺失值，这种相似通常通过距离的定义实现。

热卡填充：通过将缺失值与数据集中其他几个关键变量（具有完整值）的值进行匹配来处理缺失值。

MICE：一种基于蒙特卡洛马尔可夫链(MCMC)的插补方法，该方法对每个缺失数据插补得到多个数据集，并将多个结果进行组合分析得到最终估计值。

SFI：利用矩阵填充对缺失函数型数据进行修复。

HFI：将 SFI 中的惩罚项由核范数替换为l_0范数。

PACE(Yao 等，2005)：利用 Karhunen-Loève 展开直接表示轨迹，并从数据中学习特征函数。

FM(Li & Chiou，2021)：利用均值函数填充样本曲线中的缺失值。

NFMC：本文所提 CNFMC 方法的初始插补步骤，以检验 CNFMC 方法中聚类的有效性。

3.3.3 交通流数据相关性分析

基于皮尔逊相关系数计算不同交通流量样本之间的相关性大小，如图 3.1 所示，图中颜色越红，表示相关性越强。观察图 3.1 可知，尽管交通流量数据具有较强的正相关性，且相关系数实际分布在一个较大范围[0.4,1]内，说明不同检测器和工作日，交通流量样本的相关性波动较大。随机选取一条测试样本，并寻找与其最相关以及最不相关的样本，结果如图 3.2 所示。测试样本与最相关、最不相关样本之间的皮尔逊相关系数分别为 0.993 和 0.366，显然，最不相关样本的变化轨迹呈现更复杂的形状，说明不同样本之间的变化模式有显著差异。基于交

通流量样本之间的相关性划分不同的类别，类内样本具有相似的变化模式、相关性较大，最相似的样本对于缺失值插补能提供更可靠的信息。

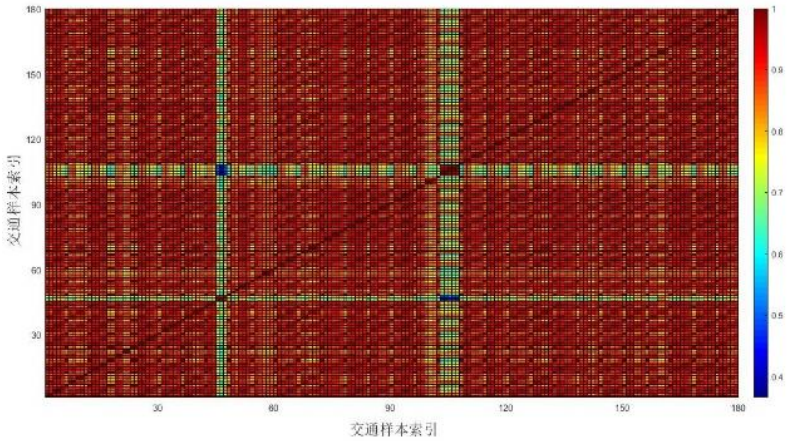


图 3.1 样本间皮尔逊相关系数矩阵

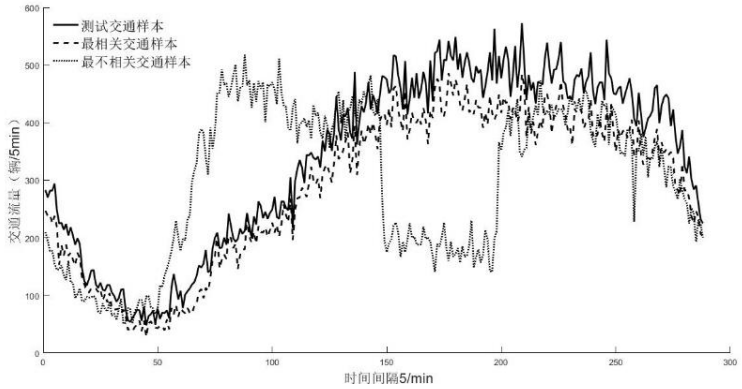


图 3.2 最相关和最不相关交通样本

3.3.4 聚类数的确定与聚类结果展示

在 CNFMC 方法中，聚类数k对缺失值的插补有显著影响。k值越小，类内样本方差增大，填充缺失值时可用相关信息较少，导致插补误差增大；反之，k值越大，聚类数越多，类内样本方差减小，且各样本相关性较强，但同时也使类内样本数减少，进而影响 NFMC 方法的插补性能。此外，为说明 CNFMC 方法中集成学习的有效性，还需研究无集成学习时(即只进行步骤 1 和步骤 2)填充方法的插补性能，故在不同缺失率下，分别比较所提方法在无集成学习与有集成学习时，插补性能随聚类数k的变化，如图 3.3 所示。观察图 3.3 可知，无集成学习时，插补误差随k的增大而减少，但当k超过某一值时，插补误差逐渐增大，表明过多

的聚类数反而会增大方法的插补误差，降低插补性能；而有集成学习时，插补误差随k的增大不断减小，集成不同的插补结果比单一聚类数下的插补误差更小，意味着自加权集成学习算法显著提高了插补精度。实验测试结果表明当$k=9$时，插补误差趋于平稳，故确定 CNFMC 方法在交通流量数据插补实验中的聚类数为 9。

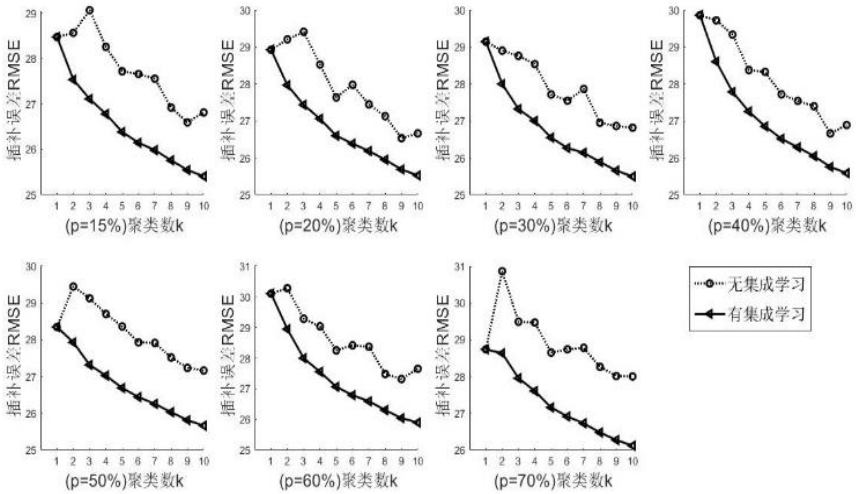


图 3.3 不同缺失率下插补误差 RMSE 随聚类数的变化曲线

为进一步说明交通流量样本之间的相关性以及存在的潜在差异，选取 9 类中的 3 类可视化展示其部分样本，如图 3.4 所示，一横排代表一类。从图 3.4 可以更加直观地看到，不同类别之间交通流量样本的变化趋势、达到车流量早晚峰时间以及持续时间等有显著差异，而同一类样本的变化模式更相似，样本轨迹基本相同。因此，聚类结果展示了预期的效果，即类间差异大、类内差异小。

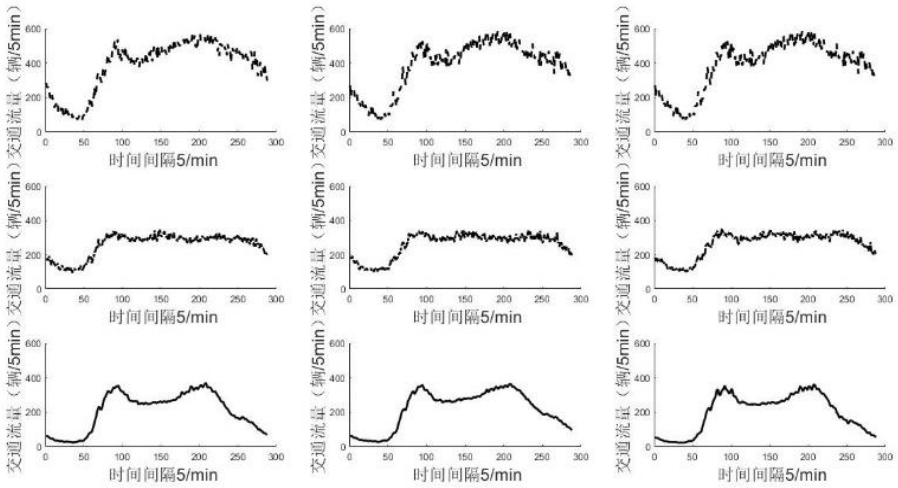


图 3.4 CNFMC 方法聚类的部分结果

3.3.5 插补结果分析

不同缺失率下，交通流量数据插补结果如表 3.1、表 3.2、表 3.3 所示，可以得到以下结论。

(1) 传统多元统计插补方法中热卡填充、MICE 的插补性能显著优于均值填充、线性插值、KNN，然而热卡填充、MICE 的插补误差受缺失率的影响较大，鲁棒性差；均值填充、线性插值在所有缺失率的情况下，RMSE 高于 K 近邻算法 4.93%~32.38%，其原因在于这两种方法没有充分考虑到交通流样本曲线之间的相关性。

(2) 函数型插补方法 SFI、HFI、FPCA、FM 以及 NFMC 方法的插补误差均低于传统多元统计插补方法，说明这 5 种方法在插补交通流函数型数据时具有一定的优越性。同时，由于 NFMC 方法融合了函数型数据分析、矩阵填充以及非负矩阵分解等思想，其 RMSE 比 SFI、HFI、FPCA、FM 低 16.90%~67.11%，插补性能在 5 种方法中最优。

(3) 不同缺失率下，CNFMC 方法的 RMSE 均显著小于其他缺失值插补方法，插补性能最优。此外，相较于 NFMC 方法，CNFMC 方法利用集成学习将多个插补结果进行融合，RMSE 显著降低了 10.75%~14.16%。因此，通过探索交通流数据的潜在变化模式，利用样本曲线之间的相关性以及差异性填充缺失值，并对不同插补结果进行集成学习，可以显著提高 CNFMC 方法的插补性能。

(4) 为了进一步比较插补性能，分别计算插补方法的 RMSE、MAE、MAPE

曲线下面积(Area Under Curve，AUC)作为度量指标(Chiou，2014)，如 $\mathrm{AUC}=\int_{p_L}^{p_U}\mathrm{RMSE}(p)dp$，本文中定义$p_L=0.15$，$p_U=0.70$。各个插补方法的 AUC 插补误差如图 3.5 所示。观察图 3.5 可知，相较于其他方法，CNFMC 方法的 AUC 误差最小，NFMC 次之，说明整体上 CNFMC 方法的插补精度优于其他方法。

(5) 对比表 3.1、表 3.2、表 3.3 发现，所有插补方法的误差 RMSE、MAE、MAPE 均随着缺失率的增多而增大，说明函数型数据出现大规模连续缺失对缺失值的插补影响较大。通过观察可以发现，CNFMC 方法的插补结果在所有缺失率下都十分稳定，RMSE、MAE、MAPE 平均变化仅为 0.47%、0.37%、0.00%，故本文所提 CNFMC 方法受缺失率的影响较低，鲁棒性较好。

(6) 不同缺失率下各个插补方法运行一次的时间消耗如表 3.4 所示。观察表 3.4 可知，虽然 CNFMC 方法相较于均值填充、线性插值、K 近邻算法、热卡填充耗时显著增加，但相比于 MICE、PACE、FM 优势显著，其中不同缺失率下经典函数型插补方法 PACE、FM 处理一次缺失值耗时为 CNFMC 方法的 40.27 倍~199.98 倍。实验表明，CNFMC 方法相比其他方法插补精度更高，且处理时间可控，故 CNFMC 方法在处理大规模缺失数据方面具有显著的优势。

表 3.1 不同缺失率下 RMSE 结果(10 次重复模拟结果均值±标准差)

方法	缺失率						
	15%	20%	30%	40%	50%	60%	70%
均值填充	112.8 ±0.31	113.09 ±0.32	113.52 ±0.36	114.23 ±0.37	114.91 ±0.38	116.50 ±0.25	118.29 ±0.21
线性插值	129.87 ±0.95	131.4 ±0.63	134.60 ±0.48	137.22 ±0.50	139.93 ±0.65	142.24 ±0.45	143.32 ±0.28
KNN	107.50 ±0.55	107.71 ±0.30	107.64 ±0.31	107.84 ±0.28	107.90 ±0.28	108.03 ±0.21	108.26 ±0.25
热卡填充	59.74 ±1.62	67.09 ±2.82	82.98 ±2.05	96.39 ±1.83	107.43 ±2.44	118.09 ±2.71	126.66 ±1.89
MICE	58.68 ±0.46	67.87 ±0.43	83.18 ±0.48	96.31 ±0.36	107.57 ±0.52	117.87 ±0.56	127.27 ±0.68
SFI	35.44 ±0.01	35.48 ±0.02	35.57 ±0.02	35.70 ±0.03	35.89 ±0.05	36.22 ±0.05	36.88 ±0.23
HFI	35.44 ±0.01	35.48 ±0.02	35.57 ±0.01	35.68 ±0.02	35.88 ±0.07	36.19 ±0.06	36.89 ±0.11
PACE	37.67 ±0.00	37.67 ±0.01	37.74 ±0.03	37.81 ±0.04	37.90 ±0.04	37.98 ±0.02	38.22 ±0.01
FM	41.62 ±0.16	47.90 ±0.27	58.67 ±0.17	67.60 ±0.13	75.97 ±0.23	82.98 ±0.03	90.08 ±0.31
NFMC	29.45 ±2.07	28.92 ±1.87	29.14 ±1.96	30.00 ±2.11	28.92 ±1.95	30.10 ±2.14	29.62 ±1.72
CNFMC	**25.54 ±0.13**	**25.68 ±0.22**	**25.65 ±0.27**	**25.75 ±0.18**	**25.81 ±0.12**	**26.04 ±0.20**	**26.27 ±0.19**

注：粗体表示比较结果为优.

表 3.2 不同缺失率下 MAE 结果(10 次重复模拟结果均值±标准差)

方法	缺失率						
	15%	20%	30%	40%	50%	60%	70%
均值填充	95.02 ±0.38	95.20 ±0.36	95.30 ±0.3	95.66 ±0.29	95.92 ±0.32	96.68 ±0.27	97.54 ±0.21
线性插值	108.43 ±0.89	109.16 ±0.62	111.0 ±0.48	112.45 ±0.45	114.01 ±0.52	115.41 ±0.39	115.66 ±0.25
KNN	90.58 ±0.52	90.87 ±0.24	90.78 ±0.25	90.85 ±0.28	90.86 ±0.23	90.86 ±0.21	90.91 ±0.21
热卡填充	17.84 ±0.59	23.09 ±1.20	34.83 ±1.21	46.79 ±1.18	58.38 ±1.76	70.35 ±2.13	81.42 ±1.40
MICE	17.60 ±0.16	23.56 ±0.16	35.35 ±0.24	47.28 ±0.20	59.01 ±0.34	70.95 ±0.38	82.78 ±0.49
SFI	26.28 ±0.01	26.31 ±0.01	26.38 ±0.01	26.45 ±0.01	26.55 ±0.04	26.74 ±0.05	27.04 ±0.05
HFI	26.29 ±0.01	26.32 ±0.01	26.38 ±0.01	26.43 ±0.02	26.55 ±0.03	26.75 ±0.03	27.08 ±0.04
PACE	28.54 ±0.05	28.49 ±0.04	28.57 ±0.05	28.65 ±0.04	28.66 ±0.02	28.81 ±0.06	29.08 ±0.09
FM	**13.60 ±0.09**	18.11 ±0.11	27.09 ±0.05	36.03 ±0.04	45.32 ±0.16	54.18 ±0.007	63.51 ±0.25
NFMC	20.78 ±1.65	20.40 ±1.45	20.52 ±1.54	21.18 ±1.75	20.45 ±1.52	21.41 ±1.73	20.81 ±1.32
CNFMC	17.54 ±0.12	**17.68 ±0.17**	**17.63 ±0.19**	**17.76 ±0.14**	**17.69 ±0.06**	**17.91 ±0.13**	**17.94 ±0.14**

表 3.3 不同缺失率下 MAPE(%)结果(10 次重复模拟结果均值±标准差)

方法	缺失率						
	15%	20%	30%	40%	50%	60%	70%
均值填充	20±0.03	21±0.04	22±0.01	23±0.02	26±0.01	27±0.02	29±0.01
线性插值	19±0.03	20±0.03	21±0.02	23±0.02	25±0.01	26±0.01	29±0.01
KNN	28±0.03	32±0.04	31±0.03	30±0.02	30±0.01	30±0.01	30±0.01
热卡填充	7±0.00	10±0.00	15±0.00	20±0.00	26±0.00	31±0.01	36±0.00
MICE	7±0.00	10±0.00	15±0.00	21±0.00	26±0.00	31±0.00	37±0.00
SFI	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00
HFI	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00	14±0.00
PACE	15±0.00	15±0.00	15±0.00	15±0.00	15±0.00	15±0.00	15±0.00
FM	**6±0.00**	8±0.00	12±0.00	16±0.00	21±0.00	25±0.00	29±0.00
NFMC	9±0.00	9±0.00	9±0.00	9±0.00	9±0.00	9±0.00	9±0.00
CNFMC	7±0.00	**7±0.00**	**7±0.00**	**7±0.00**	**7±0.00**	**7±0.00**	**7±0.00**

注：粗体表示比较结果为优.

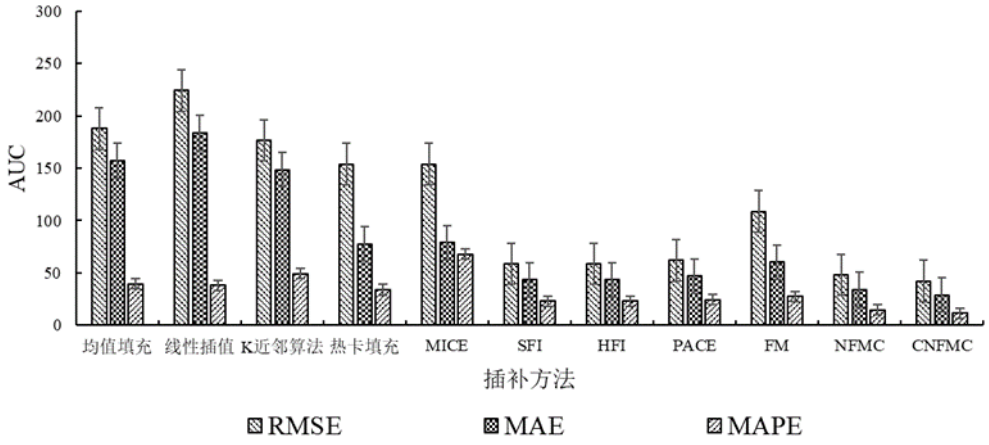


图 3.5 不同插补方法的 AUC 误差

表 3.4 不同缺失率下各插补方法运行一次的时间消耗(min)

	缺失率						
方法	15%	20%	30%	40%	50%	60%	70%
均值填充	0.0004	0.0005	0.0005	0.0006	0.0005	0.0005	0.0005
线性插值	0.0006	0.0008	0.0008	0.0007	0.0008	0.0008	0.0008
KNN	0.0067	0.0072	0.011	0.0161	0.0169	0.0201	0.0444
热卡填充	0.1575	0.5218	0.1634	0.1708	0.1801	0.2098	0.2372
MICE	11.2973	8.1414	6.9901	6.126	5.3389	4.4722	3.3117
SFI	0.4934	0.3413	0.2653	0.1778	0.2189	0.1974	0.1932
HFI	0.2342	0.1772	0.1693	0.1586	0.1378	0.118	0.1117
PACE	353.8971	298.7661	231.6300	170.8250	118.8527	77.8524	44.3310
FM	353.8971	327.9359	236.2693	183.0125	126.6384	87.4712	48.2901
NFMC	0.1159	0.0636	0.0647	0.0637	0.0657	0.0691	0.0671
CNFMC	1.7696	1.0398	1.0091	1.0071	1.0236	1.0248	1.1007

3.4 实证应用

在上述模拟实验验证 CNFMC 方法插补有效性的基础上，进一步验证该方法在其他数据集上的适用性。考虑到空气质量数据的函数特性以及各监测站点之间的相关性，本小节以空气质量数据集为例，将 CNFMC 方法针对大规模稀疏空气质量数据开展插补应用。具体地，利用 CNFMC 方法对 2021 年北京市 35 个监测站点的 6 种空气污染物($PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3)浓度数据进行缺失插补，以拓宽 CNFMC 方法的应用场景，增强普适性。

3.4.1 数据集

北京市空气质量监测网络包含 35 个监测站点，其中，站点名称、站点类型、经纬度坐标等基本信息如表 3.5 所示，依据北京市环境监测中心的位置信息可将站点分为 5 类(为描述方便起见，依次分别定义为第 1 类、第 2 类、第 3 类、第 4 类、第 5 类)。所用污染物浓度数据来自中国环境监测总站(http://www.cnemc.cn/)，选取北京市 35 个环境空气污染物监测站点 2021 年 1 月至 12 月的 6 种污染物日均数据作为研究对象，共计 365 天，理论上如果每个监测站点每种污染物的日均值都有效，则共有 76650 条数据，但由于某些因素导致数据存在缺失，缺失数据共计 18531 条，总体缺失率为 24.17%。而不同空气污染物的缺失率各不相同，PM_{10} 为 61%、CO 为 20%、O_3 为 19%、NO_2 为 17%、SO_2 为 17%、$PM_{2.5}$ 为 11%，其中 PM_{10} 缺失率最高，$PM_{2.5}$ 缺失率最低。

表 3.5 北京市空气质量监测站点基本信息

城六区					
编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标
1	东城东四 (116.417,39.929)	5	朝阳奥体中心 (116.397,39.982)	9	丰台小屯 (116.25528,39.87694)
2	东城天坛 (116.407,39.886)	6	朝阳农展馆 (116.461,39.937)	10	丰台云岗 (116.146,39.824)
3	西城官园 (116.339,39.929)	7	海淀万柳 (116.287,39.987)	11	石景山古城 (116.176,39.914)
4	西城万寿西宫 (116.352,39.878)	8	海淀四季青 (116.23052,40.03)	12	石景山老山 (116.20764,39.90886)
西北部					
编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标
1	昌平镇 (116.234,40.217)	3	定陵(对照点) (116.22,40.292)	5	延庆石河营 (116.00138,40.46327)
2	昌平南邵 (116.27603,40.21651)	4	延庆夏都 (115.972,40.453)		
东北部					
编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标
1	怀柔镇 (116.628,40.328)	4	密云新城 (116.85152,40.4088)	7	顺义新城 (116.655,40.127)
2	怀柔新城 (116.6018,40.3118)	5	平谷镇 (117.118,40.143)	8	顺义北小营 (116.6853,40.16087)
3	密云镇 (116.832,40.37)	6	平谷新城 (117.0854,40.15353)		
东南部					

编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标
1	通州永顺 (116.67503,39.93435)	3	大兴黄村 (116.404,39.718)	5	亦庄开发区 (116.506,39.795)
2	通州东关 (116.6996,39.9131)	4	大兴旧宫 (116.47456,39.78284)	6	京东南区域点 (116.78437,39.63606)
西南部					
编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标	编号	站点名称与经纬度坐标
1	门头沟双峪 (116.106,39.937)	3	房山良乡 (116.136,39.742)		
2	门头沟三家店 (116.09122,39.96926)	4	房山燕山 (115.96916,39.76419)		

注：站点信息依据 2021 年北京市空气质量监测点分类调整后的命名方式.

3.4.2 监测点数据缺失情况

如表 3.6 所示，不同污染物 35 个监测站点的缺失情况存在很大差异，以 PM_{10} 为例，缺失最严重的监测站点(京东南区域点)缺失率高达 78%，大规模缺失导致京东南区域点监测到的数据有效性不足，若直接使用该数据，可能会产生较大误差，降低统计推断的精度，最终导致统计分析结果偏误。

表 3.6 各监测站点 6 种污染物数据缺失统计表(%)

	最小值	平均值	最大值	25%分位数	50%分位数	75%分位数
$PM_{2.5}$	3.84	10.83	20.00	6.99	10.68	15.7
PM_{10}	50.00	60.86	78.00	55.00	59.00	66.50
NO_2	5.21	17.09	34.79	13.56	18.36	20.00
SO_2	4.93	16.90	35.34	14.79	18.08	19.73
O_3	9.32	19.44	56.71	15.75	17.81	21.23
CO	9.32	19.98	40.00	17.12	20.82	22.74

3.4.3 缺失机制分析

本节空气质量数据为混合 PM/IM 的缺失模式。北京市空气质量数据的 PM 和 IM 频率分布情况呈现如表 3.7 所示，每个数据集中大部分缺失模式为 PM，其中，缺失比率最高达到该数据集总缺失的 88.58%，缺失间隔长度从 2 天至 17 天不等。以 PM_{10} 数据集为例展示缺失值的分布情况，如图 3.6 所示。图 3.6 中，左图显示了 35 个站点中 PM_{10} 缺失区间的频率分布，右图为一年中各个月份的缺失率，每个月的缺失比率各不相同，尤其 1 月份缺失率最高达到 10.62%。

表 3.7 点缺失(零长度)和不同长度缺失区间的频率分布

	缺失区间长度(day)	0	2	3	4	大于 5
$PM_{2.5}$	频率	971	106	28	13	9
	比例(%)	70.16	15.32	6.07	3.76	4.70
PM_{10}	频率	791	467	307	243	332
	比例(%)	36.96	21.82	14.35	11.36	15.51
NO_2	频率	5449	207	41	18	12
	比例(%)	88.58	6.75	2.01	1.17	1.22
SO_2	频率	3921	193	44	12	13
	比例(%)	85.76	8.44	2.89	1.05	1.86
O_3	频率	1420	259	96	22	27
	比例(%)	77.85	14.20	5.26	1.21	1.48
CO	频率	1502	305	71	17	22
	比例(%)	78.35	15.91	3.7	0.89	1.15

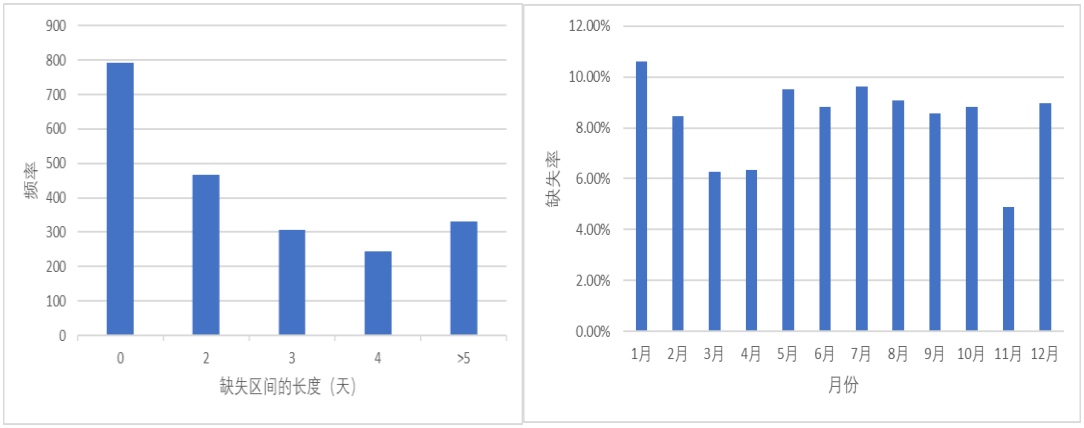


图 3.6 PM_{10} 缺失区间的频率分布(左)及每个月份的缺失比率(右)

由于所研究样本时间跨度较长，故选取缺失率最高的 1 月份进行描述统计分析，如图 3.7 所示。图 3.7 中左图的横坐标为一年中第 1 天至第 31 天(2021 年 1 月)，纵坐标为缺失数据的个数，可以看到 35 个站点一天中缺失数据的个数最多为 35 个；在中间图中，蓝色部分表示各站点每天的观测数据，红色部分表示缺失的数据，其中，缺失数据大多分布在 1 月中下旬且区间缺失尤为明显；右图利用灰度的深浅程度近似表示 PM_{10} 浓度观测数据取值的大小，颜色越浅表示值越小，颜色越深表示取值越大，红色方块默认表示数据缺失点，该图每一行为各监测站点 31 天的缺失情况，每一列为一天中 35 个站点的缺失情况，可以看出，红色方块集中表现为“条状、块状”分布，故 PM_{10} 浓度数据存在大规模“条状、块状”缺失、缺失比例较高的特点，为混合 PM/ IM 缺失模式。

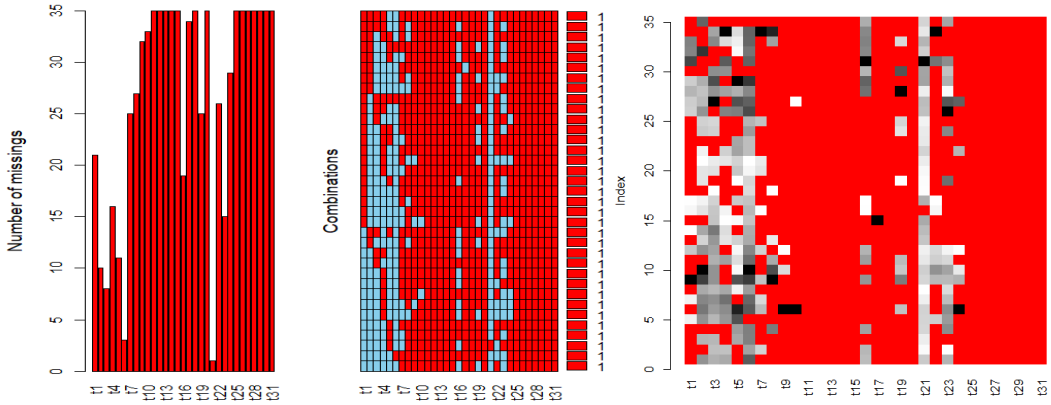


图 3.7 数据缺失数量及分布情况

3.4.4 有效性检验

本节采用 CNFMC 方法对 2021 年北京市 35 个监测站点的 6 种空气污染物数据进行了插补，为说明 CNFMC 方法插补空气质量缺失数据的准确性，计算每种污染物真实观测值与预测值之间的误差，如表 3.8 所示。由表 3.8 可以直观地看出，相比经典的函数型数据插补方法 SFI、HFI，NFMC 方法在插补空气质量缺失数据时具有较好的估计精度，由于 CNFMC 方法引入空气质量数据的类信息，使得该方法的准确性更强。

表 3.8 6 种空气污染物真实观测值与预测值之间的 RMSE、MAE

	误差指标	$PM_{2.5}$	PM_{10}	NO_2	SO_2	O_3	CO
SFI	RMSE	31.07	44.67	11.99	1.04	18.69	0.29
	MAE	20.88	27.54	8.83	0.67	14.71	0.21
HFI	RMSE	31.07	44.67	11.99	1.03	18.69	0.29
	MAE	20.88	27.54	8.83	0.67	14.71	0.21
NFMC	RMSE	26.04	39.20	11.00	0.97	16.25	0.24
	MAE	17.22	24.26	8.05	0.65	13.02	0.18
CNFMC	RMSE	**25.88**	**37.54**	**10.62**	**0.92**	**15.98**	**0.24**
	MAE	**17.14**	**23.56**	**7.90**	**0.63**	**12.81**	**0.18**

注：粗体表示比较结果为优.

为进一步验证 CNFMC 方法估计值的有效性，本文使用皮尔逊相关系数来测度缺失值和填充值之间的相关关系，二者相关性越高，拟合效果越好，反之亦然。本小节选取每种污染物中缺失率前二的站点验证 CNFMC 方法的有效性以及填充值的准确性，例如，对于 CO，分别计算亦庄开发区(缺失 40.00%)、西城万寿

西宫(缺失 29.86%)删除缺失值后的数据、填充值的数据、填充缺失值后数据的相关系数并绘制散点图，如果 3 个相关系数值比较接近，则说明插补值符合原有数据规律，CNFMC 方法有效(张波和宋国君，2022)。同理，对于 $PM_{2.5}$、SO_2、NO_2、O_3、PM_{10}，分别计算房山燕山(缺失 20.00%)与丰台云岗(缺失 17.53%)、亦庄开发区(缺失 35.34%)与西城万寿西宫(缺失 27.95%)、亦庄开发区(缺失 34.79%)与西城万寿西宫(缺失 28.77%)、房山良乡(缺失 56.71%)与亦庄开发区(缺失 40.27%)、京东南区域点(缺失 78.08%)与房山燕山(缺失 74.52%)之间三组相关系数并绘制散点图，如图 3.8 所示。从图 3.8 可以直观地看出，填充前、填充后以及填充值的 3 个相关系数之间比较接近，表明经由 CNFMC 方法估计的填充值符合各监测站点之间的相关规律，插补效果较好。

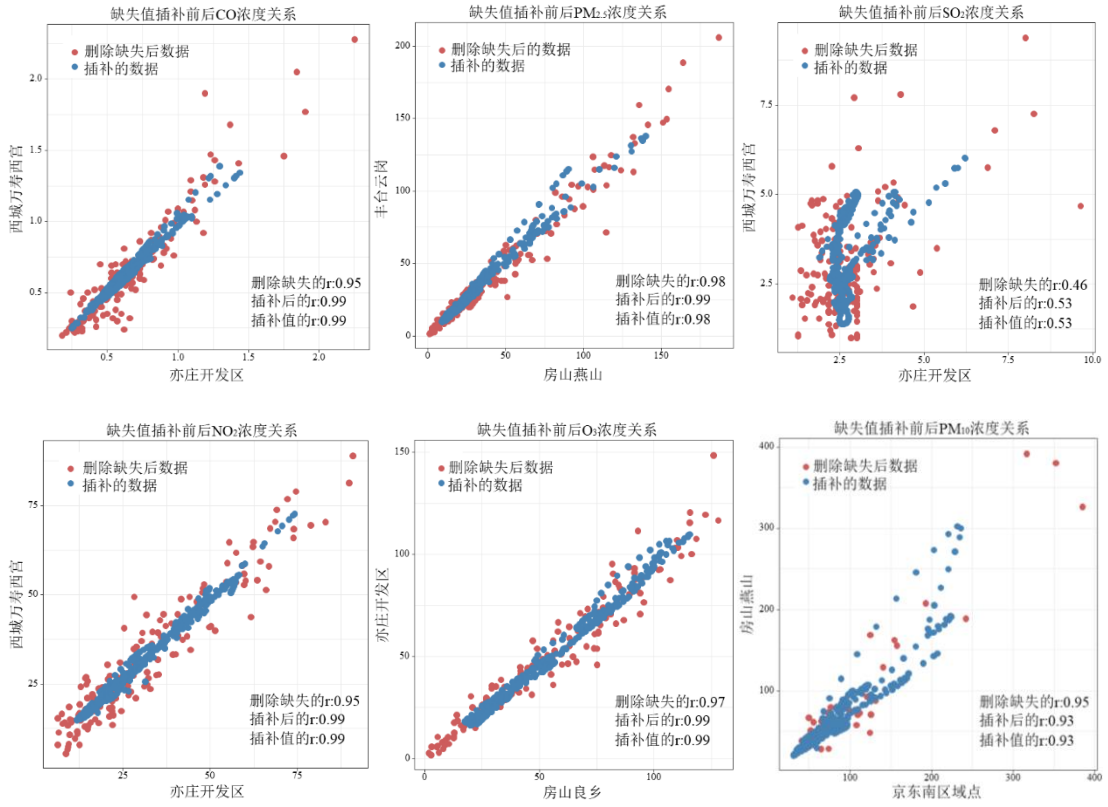


图 3.8　6 种污染物缺失值插补前后的相关性

为进一步说明 CNFMC 方法插补空气质量数据缺失值的准确性，同样选取每种污染物中缺失率最高的前两个站点，分别将对应站点污染物的真实观测值和相应预测值作为横纵坐标，给出真实值关于预测值的散点图，缺失率最高的站点如图 3.9 所示，其余见附录(图 3.10)。从图 3.9 可以直观看出，每个站点的样本点均

集中在直线附近，说明真实值与预测值之间的误差较小，即 CNFMC 方法的插补性能好，预测精度较高。

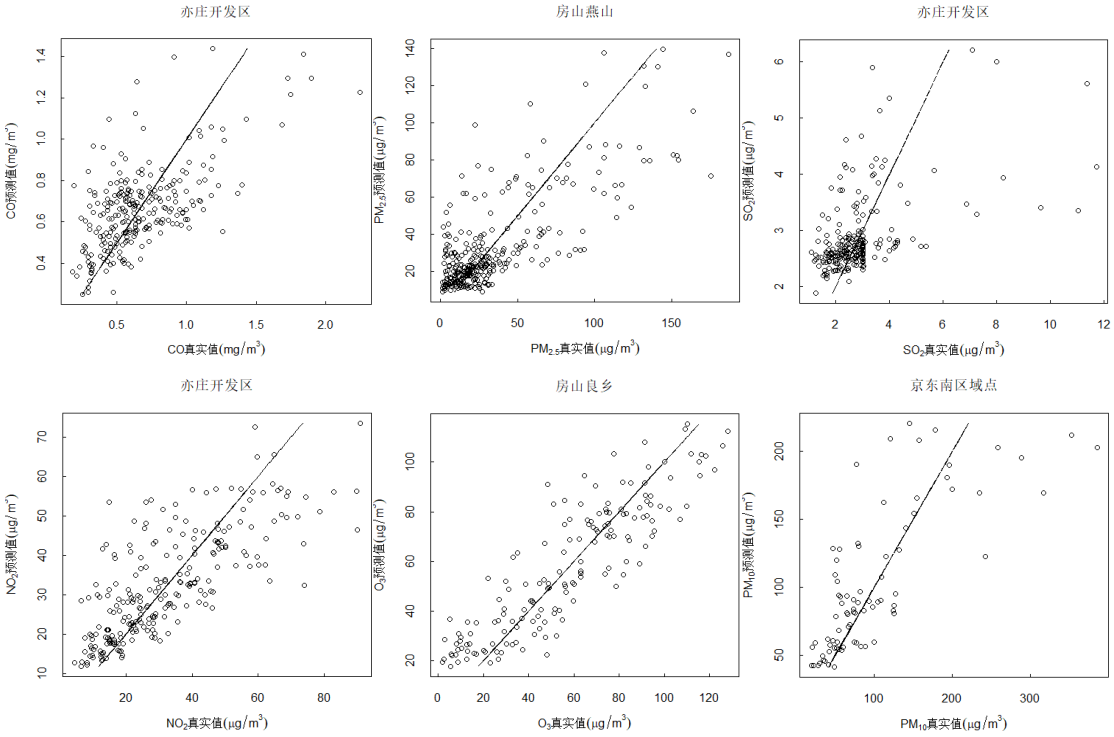


图 3.9　6 种污染物缺失率最高的站点真实观测值关于预测值的散点图

3.5 本章小结

本章首先考虑到结合类信息可以进一步提高缺失数据的插补精度，利用函数型数据聚类分析方法搜索具有相似变化模式的同质子群，由于不同的聚类数会得到不同的插补结果，应用集成学习自加权方法将不同的插补结果进行融合，提出一种融合类信息的函数型矩阵填充方法(CNFMC)；其次，采用乘性迭代方法优化求解更新规则、根据辅助函数法证明算法局部收敛性，并给出算法的复杂性分析；最后在两个不同的数据集上验证了 CNFMC 方法的准确性，相较于 KNN、MICE、PACE 等 10 种插补方法，CNFMC 方法具有显著的插补优势。

4 基于图正则化的多视角函数型矩阵填充方法

4.1 方法框架

对于多视角数据，各视角间具于高度的相关性，每个视角都包含其他视角的补充信息。通过有效利用不同视角间潜在的互补信息，可以提高多视角缺失数据的插补精度。众所周知，高独立性意味着两个视角的互补信息更多(Niu 等，2013)，而传统的独立性标准，如 Kendall 相关和 Spearman 相关只能测量线性相关性，而不能评估混合和复杂的非线性关系。相比之下，Hilbert-Schmidt 独立性准则不仅可以测量非线性依赖关系，而且可以很容易地转化为矩阵迹的形式，极大地方便了问题的优化求解。因此，本文利用 HSIC 探索各视角之间的互补信息，进一步提高缺失估计的准确性。

给定多视角矩阵$\boldsymbol{Y}=(\tilde{\boldsymbol{Y}}_1,\cdots,\tilde{\boldsymbol{Y}}_v,\cdots,\tilde{\boldsymbol{Y}}_{n_v})$，MVNFMC 方法的目标函数为

$$\min_{\boldsymbol{U}_v,\boldsymbol{V}_v\geq 0}\{\sum_{v=1}^{n_v}\theta_v||\boldsymbol{O}\odot(\tilde{\boldsymbol{Y}}_v-\boldsymbol{\Phi}\boldsymbol{U_v}\boldsymbol{V_v}^T)||_F^2+\sum_{v=1}^{n_v}\sum_{s=1}^{n_v}\theta_{vs}||\boldsymbol{V}_v-\boldsymbol{V}_s||_F^2\} \tag{4.1}$$

式(4.1)中，$\boldsymbol{\Phi}\in\mathbb{R}^{m\times r}$为基矩阵，$\boldsymbol{U}_v\in\mathbb{R}^{r\times d}$和$\boldsymbol{V}_v\in\mathbb{R}^{n\times d}$分别为第$v$个视角行、列的潜变量，$\odot$为 Hadamard 矩阵乘积。$\boldsymbol{O}\in\mathbb{R}^{m\times n}$为指示矩阵，当$\boldsymbol{Y}_{v,ij}\neq 0$时，$\boldsymbol{O}_{ij}=1$，反之为 0。式(4.1)第一项为多视角函数型矩阵填充，第二项为联合正则化，有助于综合各视角信息。

将最优图正则化项和 HSIC 策略集成到多视角函数型矩阵填充框架中，提出基于图正则化的多视角函数型矩阵填充方法(GMVNFMC)，其目标函数为

$$\begin{aligned}\mathcal{O}=&\min_{\boldsymbol{U}_v,\boldsymbol{V}_v}\sum_{v=1}^{n_v}\frac{1}{2}\theta_v||\boldsymbol{O}\odot(\tilde{\boldsymbol{Y}}_v-\boldsymbol{\Phi}\boldsymbol{U_v}\boldsymbol{V_v}^T)||_F^2+\sum_{v=1}^{n_v}\sum_{s=1}^{n_v}\theta_{vs}||\boldsymbol{V}_v-\boldsymbol{V}_s||_F^2\\&+\frac{1}{2}\mu\sum_{v=1}^{n_v}\theta_v\mathrm{tr}(\boldsymbol{V}_v{}^T\boldsymbol{L}_v\boldsymbol{V}_v)+\frac{1}{2}\beta\sum_{v\neq s}\mathrm{HSIC}\left(\boldsymbol{V}_v,\boldsymbol{V}_s\right)\\&\text{s.t. }\boldsymbol{U}_v\geq 0,\ \boldsymbol{V}_v\geq 0,\ \boldsymbol{L}_v=\alpha_1\boldsymbol{L}_{v1}+\alpha_2\boldsymbol{L}_{v2}\end{aligned} \tag{4.2}$$

式(4.2)中，第一项为多视角函数型矩阵填充的基本目标函数；第二项为联合正则化项，综合各视角的信息；第三项为最优图正则化项，挖掘各个视角的深层关联结构信息；第四项利用 HSIC 探索不同视角包含的互补信息。所提方法 GMVNFMC 可以同时借助不同视角之间的邻域关系和互补性进行数据插补，因

此在恢复丢失条目时将优于常规的矩阵填充方法。

本文使用内积核函数$\boldsymbol{K}_{\boldsymbol{V}_v}=\boldsymbol{V}_v\boldsymbol{V}_v^T$作为 HSIC 中的核函数。为简便起见，忽略 HSIC 中的常数项，即$(N-1)^{-2}$，得多样性约束的等价形式

$$\begin{aligned}\mathrm{HSIC}(\boldsymbol{V}_v,\boldsymbol{V}_s)&=\sum_{1<s<n_v,v\neq s}\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{V}_v}\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{V}_s})\\&=\sum_{1<s<n_v,v\neq s}\mathrm{tr}(\boldsymbol{V}_v{}^T\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{V}_s}\boldsymbol{H}\boldsymbol{V}_v)\\&=\sum_{1<s<n_v,v\neq s}\mathrm{tr}(\boldsymbol{V}_v{}^T\boldsymbol{K}\boldsymbol{V}_v)\end{aligned}\tag{4.3}$$

其中

$$\boldsymbol{K}=\sum_{1<s<n_v,v\neq s}\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{V}_s}\boldsymbol{H}$$

从而，目标函数式(4.2)可写为

$$\begin{aligned}\mathcal{O}=\min_{\boldsymbol{U}_v,\boldsymbol{V}_v}&\sum_{v=1}^{n_v}\frac{1}{2}\theta_v||\boldsymbol{O}\odot(\tilde{\boldsymbol{Y}}_v-\boldsymbol{\Phi}\boldsymbol{U}_{\boldsymbol{v}}\boldsymbol{V}_{\boldsymbol{v}}^T)||_F^2+\sum_{v=1}^{n_v}\sum_{s=1}^{n_v}\theta_{vs}||\boldsymbol{V}_v-\boldsymbol{V}_s||_F^2\\&+\frac{1}{2}\mu\sum_{v=1}^{n_v}\theta_v\mathrm{tr}(\boldsymbol{V}_v{}^T\boldsymbol{L}_v\boldsymbol{V}_v)+\frac{1}{2}\beta\sum_{v=1}^{n_v}\theta_v\mathrm{tr}(\boldsymbol{V}_v^T\boldsymbol{K}\boldsymbol{V}_v)\\&\text{s.t. }\boldsymbol{U}_v\geq 0,\ \boldsymbol{V}_v\geq 0,\ \boldsymbol{L}_v=\alpha_1\boldsymbol{L}_{v1}+\alpha_2\boldsymbol{L}_{v2}\end{aligned}\tag{4.4}$$

4.2 求解算法

4.2.1 求解过程

由于式(4.4)同时对$\boldsymbol{U}_v$和$\boldsymbol{V}_v$是非凸函数，难以获得全局最优解，因此，采用乘性迭代方法并利用 KKT 条件施加非负约束，获得一种局部最优解的交替迭代算法。

式(4.4)的增广拉格朗日函数为

$$\begin{aligned}\mathcal{L}=&\sum_{v=1}^{n_v}\frac{1}{2}\theta_v||\boldsymbol{O}\odot(\tilde{\boldsymbol{Y}}_v-\boldsymbol{\Phi}\boldsymbol{U}_{\boldsymbol{v}}\boldsymbol{V}_{\boldsymbol{v}}^T)||_F^2+\sum_{v=1}^{n_v}\sum_{s=1}^{n_v}\theta_{vs}||\boldsymbol{V}_v-\boldsymbol{V}_s||_F^2\\&+\frac{1}{2}\mu\theta_v\sum_{v=1}^{n_v}\mathrm{tr}(\boldsymbol{V}_v{}^T\boldsymbol{L}_v\boldsymbol{V}_v)+\frac{1}{2}\beta\theta_v\sum_{v=1}^{n_v}\mathrm{tr}(\boldsymbol{V}_v^T\boldsymbol{K}\boldsymbol{V}_v)\\&-\mathrm{tr}(\boldsymbol{\Lambda}_v\boldsymbol{U}_v^T)-\mathrm{tr}(\boldsymbol{\Gamma}_v\boldsymbol{V}_v^T)\end{aligned}\tag{4.5}$$

其中$\boldsymbol{\Lambda}_v$，$\boldsymbol{\Gamma}_v$为拉格朗日乘子矩阵。依次更新求解$\boldsymbol{U}_v$和$\boldsymbol{V}_v$，具体求解更新规则如下。

(1) 固定$\boldsymbol{V}_v$，更新$\boldsymbol{U}_v$。

对式(4.5)关于$\boldsymbol{U}_v$求偏导，并令$\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}_v}=0$，则有

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}_v}&=\frac{\frac{1}{2}\theta_v\partial\ \mathrm{tr}\left(\boldsymbol{U}_v^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v-2\boldsymbol{U}_v^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v\right)-\partial\ \mathrm{tr}\ (\boldsymbol{\Lambda}_v\boldsymbol{U}_v^T)}{\partial \boldsymbol{U}_v}\\&=\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v-\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v-\boldsymbol{\Lambda}_v\\&=0\end{aligned}$$

从而可得

$$\boldsymbol{\Lambda}_v=\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v-\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v \tag{4.6}$$

非负约束使得 KKT 条件$\boldsymbol{\Lambda}_v\odot\boldsymbol{U}_v=0$成立，即满足

$$[\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v-\theta_v\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v]_{ij}\odot\boldsymbol{U}_{vij}=0 \tag{4.7}$$

得$\boldsymbol{U}_v$的更新规则为

$$\boldsymbol{U}_{vij}\leftarrow\boldsymbol{U}_{vij}\sqrt{\frac{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v)_{ij}}{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v)_{ij}}} \tag{4.8}$$

(2) 固定$\boldsymbol{U}_v$，更新$\boldsymbol{V}_v$。

式(4.5)关于$\boldsymbol{V}_v$求偏导，并令$\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}_v}=0$，有

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}_v}&=\frac{\frac{1}{2}\theta_v\partial\ \mathrm{tr}\left(\boldsymbol{V_v}\boldsymbol{U}_{\boldsymbol{v}}^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)-2\boldsymbol{\Phi}\boldsymbol{U_v}\boldsymbol{V}_{\boldsymbol{v}}^T(\boldsymbol{O}^T\odot\boldsymbol{O}^T\odot{\tilde{\boldsymbol{Y}}_v}^T)\right)}{\partial \boldsymbol{V}_v}\\&\quad+\frac{\partial\ \mathrm{tr}\left(\sum_{s=1}^{n_v}\theta_{vs}(\boldsymbol{V}_v-\boldsymbol{V}_s)^T(\boldsymbol{V}_v-\boldsymbol{V}_s)+\frac{\mu}{2}\theta_v\boldsymbol{V}_v^T\boldsymbol{L}_v\boldsymbol{V}_v+\frac{\beta}{2}\theta_v\boldsymbol{V}_v^T\boldsymbol{K}\boldsymbol{V}_v-\boldsymbol{\Gamma}_v\boldsymbol{V}_v^T\right)}{\partial \boldsymbol{V}_v}\\&=\theta_v(\boldsymbol{O}^T\odot\boldsymbol{V_v}\boldsymbol{U}_{\boldsymbol{v}}^T\boldsymbol{\Phi}^T)\boldsymbol{\Phi}\boldsymbol{U}_v-\theta_v(\boldsymbol{O}^T\odot\tilde{\boldsymbol{Y}}_v^T)\boldsymbol{\Phi}\boldsymbol{U_v}+\sum_{s=1}^{n_v}\theta_{vs}(\boldsymbol{V}_v-\boldsymbol{V}_s)\\&\quad+\mu\theta_v\boldsymbol{L}_v\boldsymbol{V}_V+\beta\theta_v\boldsymbol{K}\boldsymbol{V}_v-\boldsymbol{\Gamma}_v\\&=\theta_v(\boldsymbol{O}^T\odot\boldsymbol{V_v}\boldsymbol{U}_{\boldsymbol{v}}^T\boldsymbol{\Phi}^T)\boldsymbol{\Phi}\boldsymbol{U}_v-\theta_v(\boldsymbol{O}^T\odot\tilde{\boldsymbol{Y}}_v^T)\boldsymbol{\Phi}\boldsymbol{U_v}+\sum_{s=1}^{n_v}\theta_{vs}(\boldsymbol{V}_v-\boldsymbol{V}_s)\\&\quad+\mu\theta_v(\alpha_1\boldsymbol{L}_{v1}+\alpha_2\boldsymbol{L}_{v2})\boldsymbol{V}_v+\beta\theta_v\boldsymbol{K}\boldsymbol{V}_v-\boldsymbol{\Gamma}_v\\&=0\end{aligned}$$

其中

$$\boldsymbol{W}_v=\alpha_1\boldsymbol{W}_{v1}+\alpha_2\boldsymbol{W}_{v2},\quad\boldsymbol{D}_v=\alpha_1\boldsymbol{D}_{v1}+\alpha_2\boldsymbol{D}_{v2}$$

可得$\boldsymbol{V}_v$的更新规则为

$$\boldsymbol{V}_{vij}\leftarrow\boldsymbol{V}_{vij}\sqrt{\frac{(\theta_v(\boldsymbol{O}^T\odot\tilde{\boldsymbol{Y}}_v^T)\boldsymbol{\Phi}\boldsymbol{U}_v-\mu\theta_v\boldsymbol{A}_v\boldsymbol{V}_v)_{ij}}{(\theta_v(\boldsymbol{O}^T\odot\boldsymbol{V}_v\boldsymbol{U}_v^T\boldsymbol{\Phi}^T)\boldsymbol{\Phi}\boldsymbol{U}_v+\sum_{s=1}^{n_v}\theta_{vs}(\boldsymbol{V}_v-\boldsymbol{V}_s)+\mu\theta_v\boldsymbol{D}_v\boldsymbol{V}_v+\beta\theta_v\boldsymbol{K}\boldsymbol{V}_v)_{ij}}} \tag{4.9}$$

GMVNFMC 方法的具体执行步骤如算法 4.1 所示。

算法 4.1 GMVNFMC 方法

输入：每一个视角的数据矩阵$\boldsymbol{Y}=\{\tilde{\boldsymbol{Y}}_1,\cdots,\tilde{\boldsymbol{Y}}_v,\cdots,\tilde{\boldsymbol{Y}}_{n_v}\}$，基矩阵$\boldsymbol{\Phi}$，参数$\theta_v$，$\theta_{vs}$，$\mu$，$\beta$ $(v,s=1,\cdots,n_v)$，r，$\varepsilon_0=10^{-7}$，$\varepsilon_1=10^{-4}$，最大更新迭代次数$=1000$。

过程：

1: for $v=1$ to n_v do

2： 初始化：随机生成基矩阵$\boldsymbol{U}_v^0$和系数矩阵$\boldsymbol{V}_v^0$；

3： **for** $t=1,2,\cdots$最大更新迭代次数

4： 保持$\boldsymbol{V}_v$固定，根据式(4.8)更新$\boldsymbol{U}_v^{t+1}$

5： 保持$\boldsymbol{U}_v$固定，根据式(4.9)更新$\boldsymbol{V}_v^{t+1}$

6： **if** $\left\|\boldsymbol{U}_v^{t+1}\boldsymbol{V}_v^{t+1^T}-\boldsymbol{U}_v^t\boldsymbol{V}_v^{t^T}\right\|_F^2 / \left\|\boldsymbol{U}_v^t\boldsymbol{V}_v^{t^T}\right\|_F^2 \le \varepsilon_0$ and $|\mathcal{O}_{t+1}-\mathcal{O}_t|/\mathcal{O}_t \le \epsilon_1$

7： **break**

8： **end if**

9： **end for**

10 :**end for**

输出：基矩阵$\{\boldsymbol{U}_1^{t+1},\cdots,\boldsymbol{U}_{n_v}^{t+1}\}$和系数矩阵$\{\boldsymbol{V}_1^{t+1},\cdots,\boldsymbol{V}_{n_v}^{t+1}\}$。

4.2.2 收敛性证明

由于目标函数式(4.4)无法保证获得全局最优解，可以证明算法 4.1 是局部收敛性。

定理 4.1 *目标函数式(4.4)分别在更新规则式(4.8)和式(4.9)下是单调递减的。*

证明 采用标准辅助函数法证明更新规则式(4.8)的收敛性。目标函数式(4.4)中，剔除无关项，保留与$\boldsymbol{U}_v$有关的项，有

$$\mathcal{L}(\boldsymbol{U_v}):=\mathrm{tr}\left(\boldsymbol{U}_v^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v-2\boldsymbol{U}_v^T\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v\right)$$

构造$\mathcal{L}(\boldsymbol{U_v})$的辅助函数$\boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)$

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)=&\sum_{i,j}\frac{\left(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v^t\boldsymbol{V}_v^T)\boldsymbol{V}_v\right)_{ij}\boldsymbol{U}_{vij}^2}{\boldsymbol{U}_{vij}^t}\\&-2\sum_{i,j}\left(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v\right)_{ij}\boldsymbol{U}_{vij}^t\left(1+\log\frac{\boldsymbol{U}_{vij}}{\boldsymbol{U}_{vij}^t}\right)\end{aligned}\tag{4.10}$$

则满足条件

$$\boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v)=\boldsymbol{L}(\boldsymbol{U}_v),\quad \boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)\ge\boldsymbol{L}(\boldsymbol{U}_v)$$

如果取$\boldsymbol{U}_v^{t+1}$使得

$$\boldsymbol{U}_v^{t+1}=\arg\min_t\boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)\tag{4.11}$$

成立，易知$\mathcal{L}(\boldsymbol{U}_v)$是单调递减的，即

$$\mathcal{L}(\boldsymbol{U}_v^{t+1})\le\boldsymbol{G}(\boldsymbol{U}_v^{t+1},\boldsymbol{U}_v^t)\le\boldsymbol{G}(\boldsymbol{U}_v^t,\boldsymbol{U}_v^t)\le\mathcal{L}(\boldsymbol{U}_v^t)$$

根据式(4.11)使辅助函数式(4.10)达到最小，求解$\boldsymbol{U}_v^{t+1}$。对式(4.10)关于$\boldsymbol{U}_{vij}$求偏导，得

$$\frac{\partial \boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)}{\partial \boldsymbol{U}_{vij}}=2\frac{\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v^t\boldsymbol{V}_v^T)\boldsymbol{V}_v\Big)_{ij}\boldsymbol{U}_{vij}}{\boldsymbol{U}_{vij}^t}-2\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v\Big)_{ij}\frac{\boldsymbol{U}_{vij}^t}{\boldsymbol{U}_{vij}} \tag{4.12}$$

令$\frac{\partial \boldsymbol{G}(\boldsymbol{U}_v,\boldsymbol{U}_v^t)}{\partial \boldsymbol{U}_{vij}}=0$，则有

$$\Big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v\Big)_{ij}\frac{\boldsymbol{U}_{v,ij}^t}{\boldsymbol{U}_{vij}^{t+1}}=\Big(\frac{\big(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v^t\boldsymbol{V}_v^T)\boldsymbol{V}_v\big)_{ij}}{\boldsymbol{U}_{vij}^t}\Big)\boldsymbol{U}_{vij}^{t+1}$$

从而

$$\boldsymbol{U}_{vij}^{t+1}=\boldsymbol{U}_{vij}^t\sqrt{\frac{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\tilde{\boldsymbol{Y}}_v)\boldsymbol{V}_v)_{ij}}{(\boldsymbol{\Phi}^T(\boldsymbol{O}\odot\boldsymbol{\Phi}\boldsymbol{U}_v^t\boldsymbol{V}_v^T)\boldsymbol{V}_v)_{ij}}}$$

上式即为$\boldsymbol{U}_v$的更新规则式(4.8)。类似地，可证得$\boldsymbol{V}_v$的更新规则式式(4.9)。

4.2.3 计算复杂度分析

GMVNFMC方法的时间复杂度主要体现在$\boldsymbol{U}_v$和$\boldsymbol{V}_v$的更新迭代中。对于$\boldsymbol{U}_v$的更新，需要的加法运算和乘法运算分别为$O(2m_vnr+2nrd+m_vrd+m_vnd)$、$O(2m_vn+2m_vnr+2nrd+m_vrd+m_vnd+rd)$，除法运算$O(rd)$，则在一次迭代中，更新$\boldsymbol{U}_v$的时间复杂度为$O(2m_vn+4m_vnr+4nrd+2m_vrd+2m_vnd+2rd)$。同理，更新$\boldsymbol{V}_v$的时间复杂度为$O(2m_vn+6m_vnr+6nrd+6n^2d+2n_vnd+6nd)$。故GMVNFMC方法迭代一次的时间复杂度为$O(m_vn+m_vnr)$，当迭代次数为$t$时，时间复杂度为$O(t(m_vn+m_vnr))$。因此，对于更新$n_v$个视角，GMVNFMC方法在$t$次迭代过程中总的时间复杂度为$O(tn_v(m_vn+m_vnr))$。

4.3 模拟实验

4.3.1 数据集

以不完整多视角空气质量数据集为例，通过缺失插补实验，验证 GMVNFMC 方法的有效性。本节实验选取空气质量在线监测平台[①]发布的 2014-2021 年京津

[①] http://www.aqistudy.cn/historydata. 该站点的月度数据是根据当天环保总站每小时数据计算求平均的结果，存在丢数据场景，但不影响本文的缺失值插补实验。

冀地区 14 个城市[①]的 6 种污染物完整月度数据，视角 1~6 依次分别为 CO、$PM_{2.5}$、SO_2、NO_2、O_3、PM_{10}，每个视角的数据曲线如图 4.1 所示。

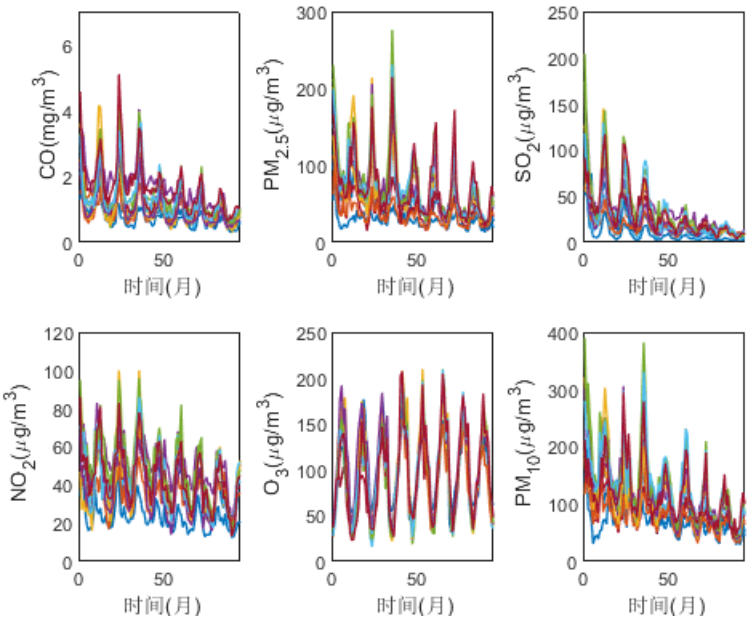


图 4.1　京津冀地区 14 个城市 6 种污染物的数据曲线

对于多视角数据集而言，缺失值的插补依赖于不同视角间的相关性。为了更深入地了解这种相关性，计算不同视角之间的皮尔逊相关系数，相关系数的绝对值越大，二者相关性越高。6 个视角之间的相关系数如表 4.1 所示。

表 4.1　不同视角间的相关系数

	视角 1	视角 2	视角 3	视角 4	视角 5	视角 6
视角 1	1	0.8261	0.7971	0.7995	-0.5380	0.7811
视角 2	0.8261	1	0.7208	0.7839	-0.5185	0.9401
视角 3	0.7971	0.7208	1	0.6399	-0.4032	0.7711
视角 4	0.7995	0.7839	0.6399	1	-0.6461	0.7701
视角 5	-0.5380	-0.5185	-0.4032	-0.6461	1	-0.4477
视角 6	0.7811	0.9401	0.7711	0.7701	-0.4477	1

表 4.1 表明，6 个视角之间存在高相关性。例如，视角 2 和视角 1、视角 6 高度正相关，相关系数分别达到 0.8261 和 0.9401；视角 4 与视角 5 负相关，相关系数为-0.6461。不同视角间的相关性保证了利用多视角数据间的互补性进行缺失插补的合理性。

[①] 北京、天津、保定、唐山、石家庄、廊坊、秦皇岛、张家口、承德、沧州、衡水、邢台、邯郸、安阳。

4.3.2 实验设置

实验主要通过 R 4.1.3 实现，实验的计算机环境为：Intel(R) Core(TM) i5-5200U CPU2.20 GHz，内存 6GB，Windows10 64 位操作系统。与第 3.3.2 节类似，首先，人为生成带有点缺失(PM)和区间缺失(IM)的条目。具体工作如下：

(1) PM：以逐点方式对缺失的条目进行统一随机采样。

(2) IM：生成不同大小的缺失区间，这些区间在每个曲线轨迹中都有均匀分布的起始位置。缺失区间的长度遵循均匀分布。

本文采用混合 PM/IM 模式，其中混合 PM/IM 模式将 PM 和 IM 结合用于单独的曲线轨迹；其次为每种模式生成nmp缺失点，其中n是曲线轨迹数，m是时间点数，p是缺失率并分别设定为 30%、40%、50%、60%和 70%。其次，在不同缺失比例设定下，将 GMVNFMC 方法与均值填充、线性插值、KNN、SFI、HFI 以及 MVNFMC 方法进行对比，并利用 RMSE、NRMSE 评估每种方法的插补性能。

4.3.3 不同缺失率的消融实验

为说明 GMVNFMC 方法的有效性，通过消融实验验证目标函数式(4.4)中正则化项(即最优图正则化、HSIC)的影响。现考虑 GMVNFMC 方法的两种特殊情况：

(1) GMVNFMC1：令目标函数式(4.4)中$\mu = 0$消除 GMVNFMC 方法中最优图正则化项的效果；

(2) GMVNFMC2：令目标函数式(4.4)中$\beta = 0$，则在估计缺失数据时不考虑各视角间的互补性；

利用 GMVNFMC1、GMVNFMC2 以及 GMVNFMC 方法修复不同缺失率的 6 种污染物数据，插补的误差评价指标结果如表 4.2 所示，可以得出如下结论：

(1) 从全局角度来看，随着缺失率的增大，GMVNFMC1、GMVNFMC2 以及 GMVNFMC 方法的插补误差反而降低，则多视角插补方法相比单一插补方法，插补性能有所提高；

(2) 表 4.2 中，当缺失率为 30%、40%、50%以及 60%时，相比于 GMVNFMC1、

GMVNFMC2，6 个视角中 GMVNFMC 方法的插补误差 RMSE 分别平均降低了 62.37%、82.85%、61.44%、58.11%以及 49.44%。故缺失率为 30%~60%时，GMVNFMC 方法的插补误差均显著小于 GMVNFMC1、GMVNFMC2，插补精度高。

(3) 缺失率为 70%时，GMVNFMC 方法的插补误插仅在极个别视角中高于 GMVNFMC1、GMVNFMC2。绘制 GMVNFMC1、GMVNFMC2 以及 GMVNFMC 方法在各个视角中插补误差的平均水平，如图 4.2、图 4.3 所示，可以直观地看出，6 个视角中 GMVNFMC 方法的插补误差均小于 GMVNFMC1 和 GMVNFMC2，表明 GMVNFMC 方法中最优图正则化以及 HSIC 有助于提升插补精度。

表 4.2 不同缺失率下各视角的 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

缺失率	对比方法	评价指标	视角 1	视角 2	视角 3	视角 4	视角 5	视角 6
30%	GMVNFMC1	RMSE	1.55	5.84	2.81	3.43	9.46	10.58
		NRMSE	1.30	0.10	0.11	0.08	0.10	0.10
	GMVNFMC2	RMSE	0.65	25.05	12.61	12.55	36.44	41.33
		NRMSE	0.55	0.41	0.50	0.30	0.37	0.39
	GMVNFMC	RMSE	**0.12**	**4.17**	**2.13**	**2.09**	**6.61**	**6.66**
		NRMSE	**0.10**	**0.07**	**0.08**	**0.05**	**0.07**	**0.06**
40%	GMVNFMC1	RMSE	1.31	6.14	2.61	3.32	9.24	10.52
		NRMSE	1.10	0.10	0.10	0.08	0.09	0.10
	GMVNFMC2	RMSE	0.70	24.68	12.81	13.42	35.78	41.72
		NRMSE	0.59	0.40	0.51	0.32	0.37	0.39
	GMVNFMC	RMSE	**0.10**	**4.46**	**2.08**	**2.20**	**6.41**	**7.08**
		NRMSE	**0.09**	**0.07**	**0.08**	**0.05**	**0.07**	**0.07**
50%	GMVNFMC1	RMSE	1.38	6.07	2.79	3.36	9.37	10.55
		NRMSE	1.16	0.10	0.11	0.08	0.10	0.10
	GMVNFMC2	RMSE	0.45	16.47	8.26	8.68	23.91	28.67
		NRMSE	0.38	0.27	0.33	0.21	0.25	0.27
	GMVNFMC	RMSE	**0.12**	**3.98**	**2.17**	**2.17**	**6.14**	**7.15**
		NRMSE	**0.10**	**0.06**	**0.09**	**0.05**	**0.06**	**0.07**
60%	GMVNFMC1	RMSE	0.25	8.33	4.54	4.43	12.18	13.69
		NRMSE	0.21	0.14	0.18	0.11	0.12	0.13
	GMVNFMC2	RMSE	0.23	8.23	4.13	4.39	12.33	13.05
		NRMSE	0.19	0.13	0.16	0.10	0.13	0.12
	GMVNFMC	RMSE	**0.11**	**4.63**	**2.09**	**2.24**	**6.25**	**6.87**
		NRMSE	**0.09**	**0.08**	**0.08**	**0.05**	**0.06**	**0.06**
70%	GMVNFMC1	RMSE	**0.11**	**3.83**	2.21	2.09	6.34	7.50
		NRMSE	**0.09**	**0.06**	0.09	0.05	0.07	0.07
	GMVNFMC2	RMSE	0.27	8.50	4.42	4.85	12.18	14.17
		NRMSE	0.22	0.14	0.18	0.12	0.12	0.13
	GMVNFMC	RMSE	0.12	4.37	**1.85**	**1.96**	**5.43**	**6.95**
		NRMSE	0.10	0.07	**0.07**	**0.05**	**0.06**	**0.07**

注：粗体表示比较结果为优.

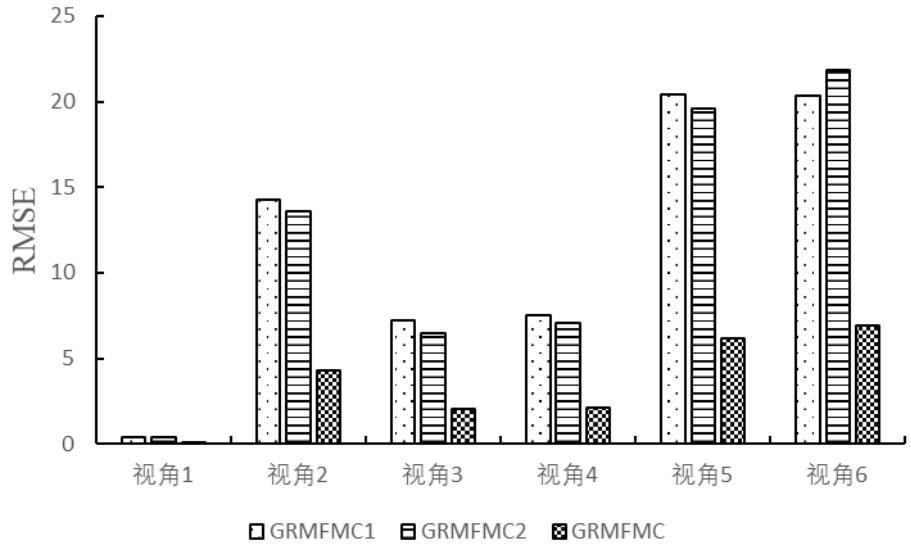


图 4.2 消融实验中三种方法的平均插补误差 RMSE

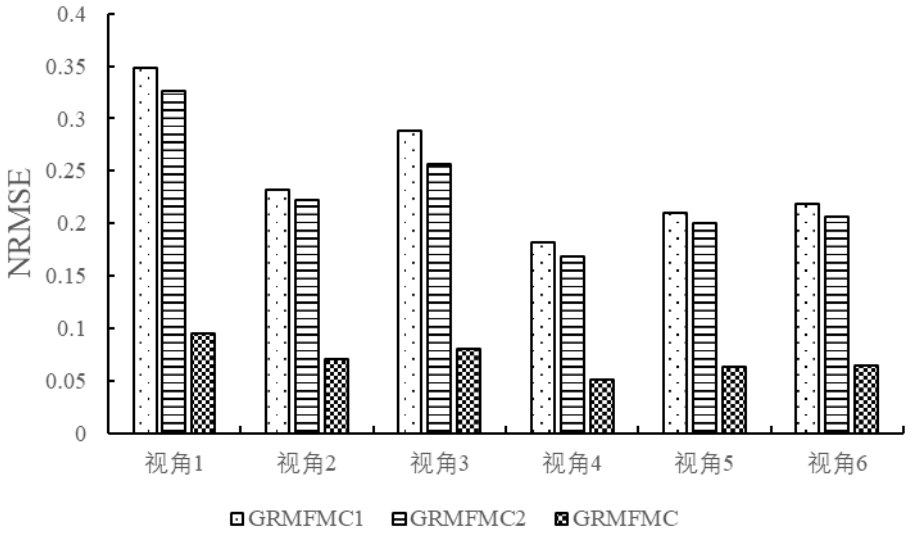


图 4.3 消融实验中三种方法的平均插补误差 NRMSE

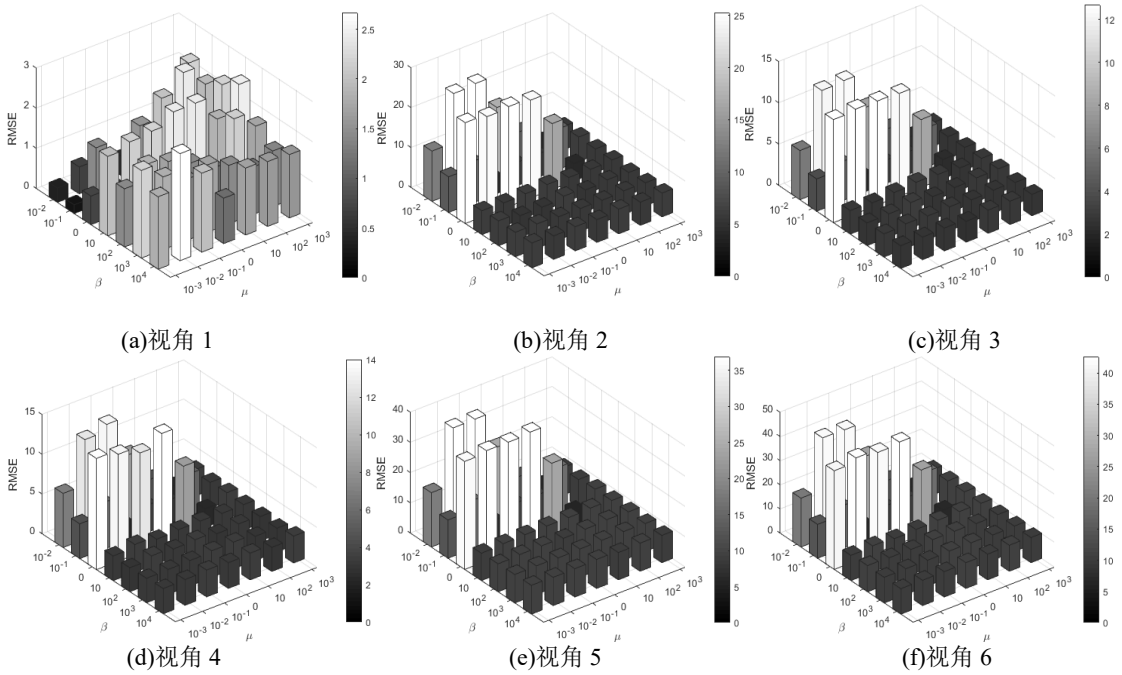
4.3.4 参数灵敏度

GMVNFMC 方法中参数μ和β的选择对于插补结果至关重要，为此，研究不同缺失下 GMVNFMC 方法对参数μ、β的敏感性。当μ、β取不同值时该方法的误差评估指标 RMSE 结果如图 4.4~图 4.8 所示。

(1) 参数μ控制最优图正则化项的权重。设置μ的取值范围为$\{10^{-3}, 10^{-2}, 10^{-1}, 0, 10, 10^{2}, 10^{3}\}$。观察图 4.4~图 4.8 可以发现：当缺失率较低时，如 30%、40%，$\mu = 0.1$的插补误差最小；当缺失率较高时，μ的取值随着缺失率的增大呈增长趋势，说明在多视角缺失数据估计中，最优图正则化项的权重随着缺失率的升高而

变大。

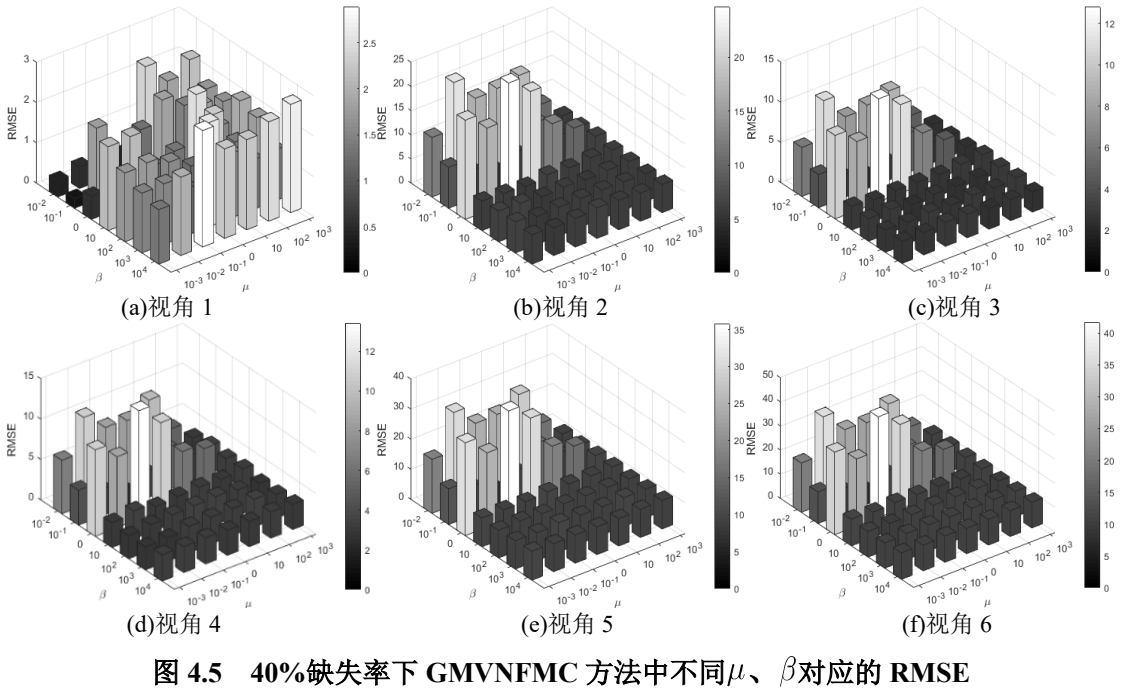
(2) 参数β控制 HSIC 项的权重。设置β的取值范围为$\{10^{-2}, 10^{-1}, 0, 10, 10^{2}, 10^{3}$ $10^{4}\}$。观察图 4.4~图 4.8 可以发现，当缺失率较低时，$\beta = 0.1$的插补效果最好；而随着缺失率的增大，β的取值降低，HSIC 项在估计缺失值时的权重变小。事实上，当缺失率增大时，各个视角内的可用样本数量减小，导致视角之间的互补信息变少。



(a)视角 1 (b)视角 2 (c)视角 3

(d)视角 4 (e)视角 5 (f)视角 6

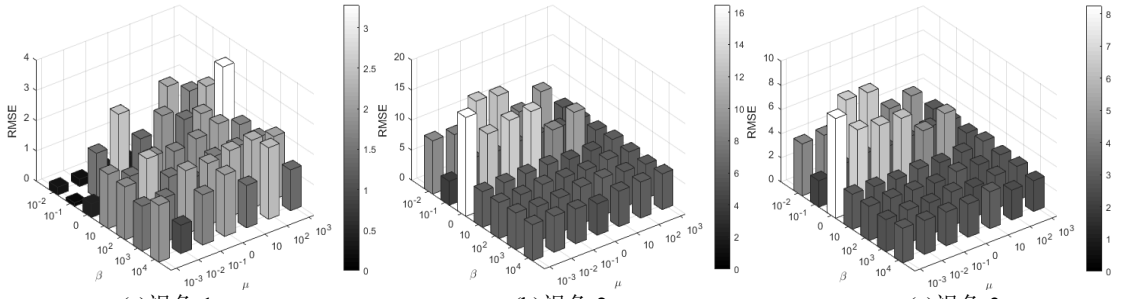
图 4.4 30%缺失率下 GMVNFMC 方法中不同μ、β对应的 RMSE



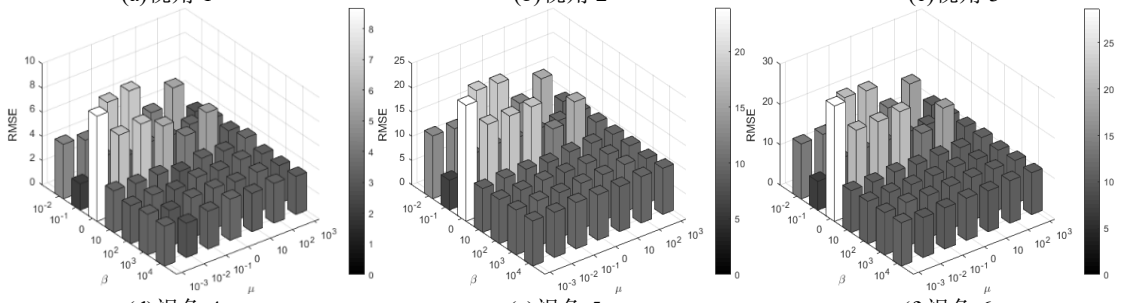
(a)视角 1 (b)视角 2 (c)视角 3

(d)视角 4 (e)视角 5 (f)视角 6

图 4.5 40%缺失率下 GMVNFMC 方法中不同μ、β对应的 RMSE

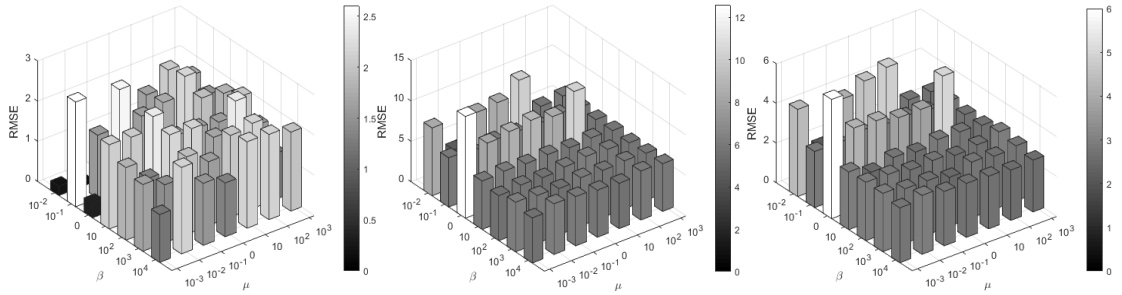


(a)视角 1 (b)视角 2 (c)视角 3

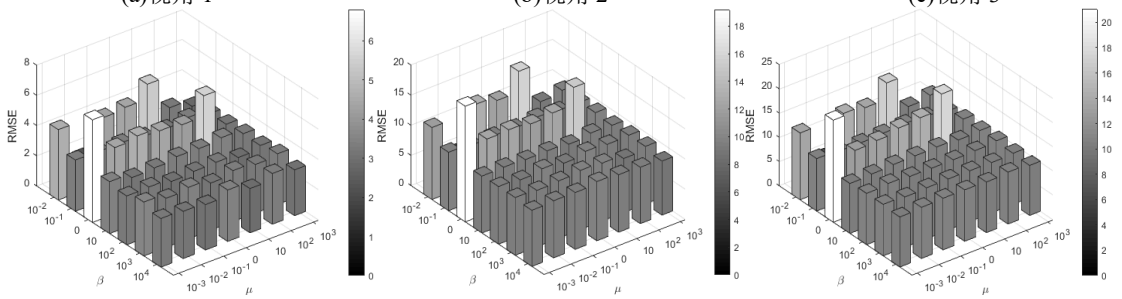


(d)视角 4 (e)视角 5 (f)视角 6

图 4.6 50%缺失率下 GMVNFMC 方法中不同μ、β对应的 RMSE

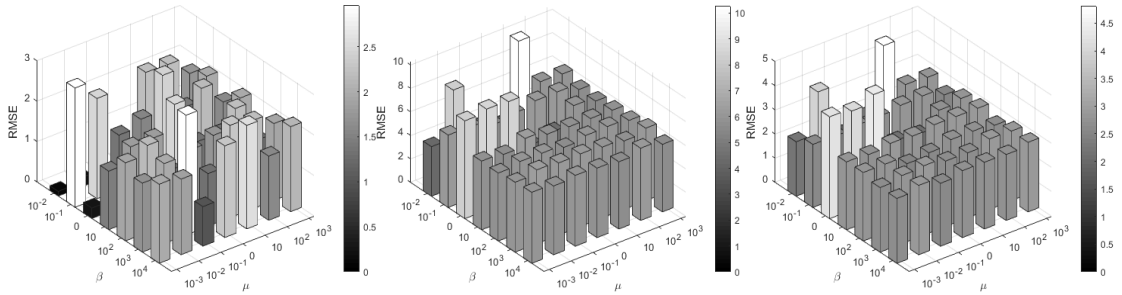


(a)视角 1 (b)视角 2 (c)视角 3

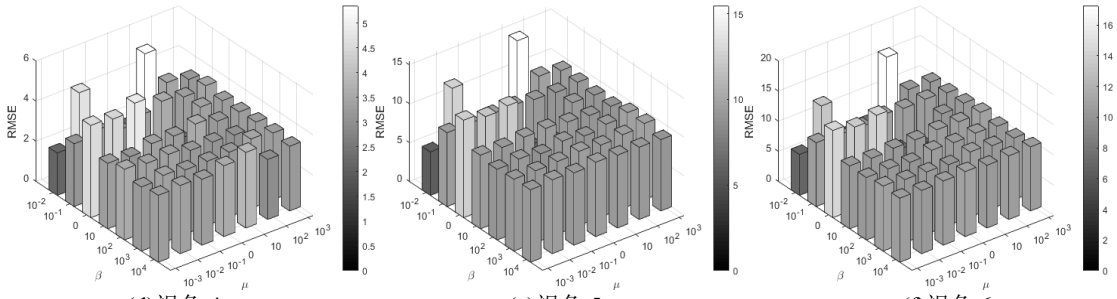


(d)视角 4 (e)视角 5 (f)视角 6

图 4.7 60%缺失率下 GMVNFMC 方法中不同μ、β对应的 RMSE



(a)视角 1 (b)视角 2 (c)视角 3



(d)视角 4 (e)视角 5 (f)视角 6

图 4.8 70%缺失率下 GMVNFMC 方法中不同μ、β对应的 RMSE

4.3.5 不同视角的缺失数据插补结果

在上述实验设计的基础上，针对不同缺失率的数据集，分别采用均值填充、线性插值、KNN、SFI、HFI、MVNFMC、以及 GMVNFMC 方法修复该数据集并进行插补性能的比较。实验结果如表 4.3~表 4.7 所示。

表 4.3 30%缺失率下 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

	视角 1		视角 2		视角 3		视角 4		视角 5		视角 6	
	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE
均值填充	0.50	0.42	24.74	0.40	12.86	0.51	10.67	0.25	**16.06**	**0.17**	37.30	0.35
线性插值	0.50	0.42	17.91	0.29	14.51	0.58	9.74	0.23	16.14	0.17	31.25	0.29
KNN	0.49	0.41	22.09	0.36	11.06	0.44	10.01	0.24	17.86	0.18	31.23	0.29
SFI	0.51	0.43	27.40	0.45	14.97	0.60	12.44	0.30	45.26	0.46	35.76	0.33
HFI	0.51	0.43	27.17	0.44	15.04	0.60	12.50	0.30	45.02	0.46	35.87	0.34
MVNFMC	6.96	5.85	138.05	2.25	34.68	1.38	23.52	0.56	30.18	0.31	28.79	0.27
GMVNFMC	**0.12**	**0.10**	**4.17**	**0.07**	**2.13**	**0.08**	**2.09**	**0.05**	**6.61**	**0.07**	**6.66**	**0.06**

注：粗体表示比较结果为优.

表 4.4 40%缺失率下 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

	视角 1		视角 2		视角 3		视角 4		视角 5		视角 6	
	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE
均值填充	0.43	0.36	23.37	0.38	14.47	0.58	11.32	0.27	14.62	0.15	36.38	0.34
线性插值	0.42	0.35	20.20	0.33	14.58	0.58	10.16	0.24	14.93	0.15	30.65	0.29
KNN	0.42	0.36	23.28	0.38	13.08	0.52	10.00	0.24	14.02	0.14	36.51	0.34
SFI	0.50	0.42	27.64	0.45	15.50	0.62	12.64	0.30	45.97	0.47	35.95	0.34
HFI	0.50	0.42	27.57	0.45	15.54	0.62	12.66	0.30	45.60	0.47	36.22	0.34
MVNFMC	7.13	5.98	134.11	2.18	33.83	1.35	23.56	0.56	31.48	0.32	28.55	0.27
GMVNFMC	**0.10**	**0.09**	**4.46**	**0.07**	**2.08**	**0.08**	**2.20**	**0.05**	**6.41**	**0.07**	**7.08**	**0.07**

注：粗体表示比较结果为优.

表 4.5 50%缺失率下 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

	视角 1		视角 2		视角 3		视角 4		视角 5		视角 6	
	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE
均值填充	0.46	0.39	26.20	0.43	14.50	0.58	11.04	0.26	17.19	0.18	37.97	0.36
线性插值	0.45	0.38	24.47	0.40	15.86	0.63	10.83	0.26	17.18	0.18	31.45	0.29
KNN	0.42	0.36	24.63	0.40	13.36	0.53	10.38	0.25	15.76	0.16	37.43	0.35
SFI	0.51	0.43	27.99	0.46	16.10	0.64	12.78	0.31	47.07	0.48	37.06	0.35
HFI	0.51	0.43	28.21	0.46	16.03	0.64	12.87	0.31	46.36	0.48	37.28	0.35
MVNFMC	6.96	5.84	139.91	2.28	34.11	1.36	24.13	0.58	29.85	0.31	28.67	0.27
GMVNFMC	**0.12**	**0.10**	**3.98**	**0.06**	**2.17**	**0.09**	**2.17**	**0.05**	**6.14**	**0.06**	**7.15**	**0.07**

注：粗体表示比较结果为优.

表 4.6 60%缺失率下 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

	视角 1		视角 2		视角 3		视角 4		视角 5		视角 6	
	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE
均值填充	0.48	0.40	26.22	0.43	15.29	0.61	11.42	0.27	15.44	0.16	40.05	0.38
线性插值	0.46	0.39	22.55	0.37	14.46	0.58	11.21	0.27	15.12	0.16	36.95	0.35
KNN	0.44	0.37	24.60	0.40	14.34	0.57	10.93	0.26	14.48	0.15	37.64	0.35
SFI	0.51	0.43	29.26	0.48	16.02	0.64	13.41	0.32	46.43	0.48	37.35	0.35
HFI	0.53	0.44	28.56	0.47	16.52	0.66	13.50	0.32	48.38	0.50	38.81	0.36
MVNFMC	7.69	6.45	134.21	2.19	34.47	1.37	24.49	0.58	31.03	0.32	30.50	0.29
GMVNFMC	**0.11**	**0.09**	**4.63**	**0.08**	**2.09**	**0.08**	**2.24**	**0.05**	**6.25**	**0.06**	**6.87**	**0.06**

注：粗体表示比较结果为优.

表 4.7 70%缺失率下 RMSE 和 NRMSE 结果(10 次重复模拟结果均值)

	视角 1		视角 2		视角 3		视角 4		视角 5		视角 6	
	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE	NRMSE
均值填充	0.51	0.43	28.25	0.46	16.12	0.64	12.30	0.29	17.28	0.18	41.54	0.39
线性插值	0.49	0.41	29.29	0.48	15.44	0.62	11.90	0.28	16.58	0.17	42.97	0.40
KNN	0.46	0.39	26.08	0.42	14.66	0.58	10.99	0.26	15.23	0.16	38.70	0.36
SFI	0.52	0.44	28.49	0.46	17.02	0.68	13.56	0.32	51.98	0.53	41.98	0.39
HFI	0.57	0.48	30.24	0.49	17.99	0.72	14.04	0.34	52.56	0.54	40.20	0.38
MVNFMC	7.23	6.07	140.98	2.30	34.20	1.36	23.33	0.56	32.42	0.33	30.35	0.28
GMVNFMC	**0.12**	**0.10**	**4.37**	**0.07**	**1.85**	**0.07**	**1.96**	**0.05**	**5.43**	**0.06**	**6.95**	**0.07**

注：粗体表示比较结果为优.

从表 4.3~表 4.7，我们可以得出如下结论：

(1) 整体上看，均值填充、线性插值、KNN、SFI 以及 HFI 插补不同视角数据的缺失值时，插补误差随着缺失率的升高而增大，然而，由于多视角插补方法 MVNFMC、GMVNFMC 在估计缺失数据时综合各个视角的信息，其插补误差随着缺失率的升高而降低。

(2) 当缺失率为 30%~70%时，GMVNFMC 方法的插补误差 RMSE 均显著小于均值填充、线性插值、KNN、SFI、HFI。例如，当缺失率为 70%时，GMVNFMC 方法在处理视角 1 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 76.47%、75.51%、73.91%、76.92%、78.95%；在处理视角 2 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 84.53%、85.08%、83.24%、84.66%、85.55%；在处理视角 3 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 88.52%、88.02%、87.38%、89.13%、89.72%；在处理视角 4 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 84.07%、83.53%、82.17%、85.55%、86.04%；在处理视角 5 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 68.58%、67.25%、64.35%、89.55%、89.67%；在处理视角 6 的缺失数据中，RMSE 相较于均值填充、线性插值、KNN、SFI、HFI 分别降低了 83.27%、83.83%、82.04%、83.44%以及 82.71%。

(3) 对比两个多视角插补方法，GMVNFMC 方法在估计缺失值时由于利用了

每条样本的二阶邻域信息以及不同视角污染物之间的多样性，6 个视角不同缺失率下其插补误差 RMSE 相较于 MVNFMC 平均降低了 98.41%、96.85%、93.98%、91.05%、80.06%以及 76.34%。此外，不同缺失率下 GMVNFMC 方法的执行时间均显著低于 MVNFMC 方法，如图 4.9 所示，说明 GMVNFMC 方法更适用于处理多视角空气质量缺失数据集。

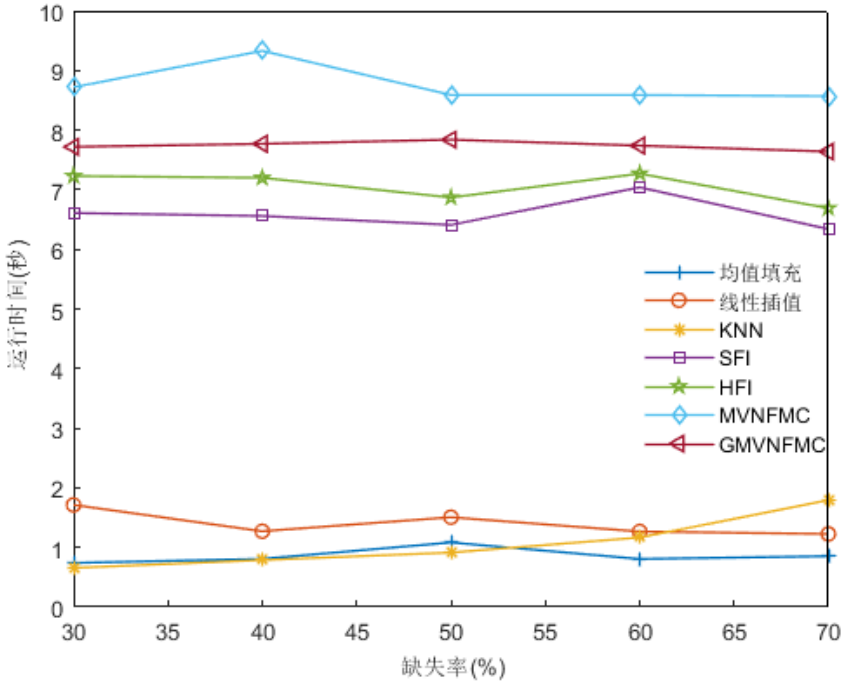


图 4.9 不同插补方法在 10 次模拟实验中的运行时间

4.4 实证应用

4.4.1 数据集

为了进一步说明 GMVNFMC 方法的实际应用效果，本节对长三角地区多视角空气质量缺失数据集进行实证应用。所用空气质量数据来自中国环境监测总站(http://www.cnemc.cn/)，选取 2021 年 1 月 1 日-2021 年 12 月 31 日长三角地区 41 个城市[①]的 6 种空气污染物以及 AQI 日均值作为研究对象，视角 1~7 分别为 CO、$PM_{2.5}$、SO_2、NO_2、O_3、PM_{10}、AQI，其描述性统计表 4.8 所示。

[①] 上海、苏州、南京、无锡、徐州、常州、南通、连云港、淮安、盐城、扬州、镇江、泰州、宿迁、杭州、宁波、温州、绍兴、湖州、嘉兴、金华、衢州、舟山、台州、丽水、合肥、芜湖、马鞍山、铜陵、池州、安庆、宣城、滁州、蚌埠、淮北、淮南、苏州、阜阳、亳州、六安、黄山

表 4.8 2021 年长三角地区不同视角空气质量数据的描述性统计

变量	数量	平均值	标准差	最小值	最大值	25%分位数	50%分位数	75%分位数
AQI	12263	54.70	32.21	8.04	440.33	34.54	48.96	66.58
$PM_{2.5}$	13596	31.45	20.95	2.38	185.33	17.04	26.15	39.75
PM_{10}	13599	59.18	44.34	4.21	996.46	32.33	49.08	74.25
SO_2	13604	7.08	2.89	1.33	35.37	5.13	6.46	8.39
NO_2	13606	27.43	14.90	2.29	113.66	16.33	24.21	35.83
CO	13604	0.63	0.20	0.15	1.98	0.49	0.60	0.74
O_3	13605	64.75	27.14	3.25	196.29	44.42	61.58	81.38

4.4.2 监测点数据缺失情况

长三角地区 2021 年的空气质量数据中，理论上如果每个城市的每种空气质量数据日均值都有效，则共有 104755 条数据，但由于某些不可预期的因素导致数据存在缺失，缺失数据共计 13038 条，总体缺失率为 12.45%。不同视角空气质量数据的缺失率各不相同，AQI 为 13.58%，$PM_{2.5}$ 为 9.15%，PM_{10} 为 9.13%，O_3 为 9.09%，CO 为 9.09%，SO_2 为 9.09%，NO_2 为 9.08%，其中 AQI 缺失率最高，NO_2 缺失率最低。同时，41 个城市的缺失情况存在较大差异，各城市 7 个视角空气质量数据的缺失统计如表 4.9 所示，其中缺失率最高的为黄山、最小为合肥。

表 4.9 各城市 7 个视角的空气质量数据缺失统计表(%)

变量	最小值	平均值	最大值	25%分位数	50%分位数	75%分位数
AQI	8.77	13.85	23.84	9.32	9.59	23.56
$PM_{2.5}$	8.49	9.15	10.41	8.77	9.04	9.32
PM_{10}	0.02	0.02	0.03	0.02	0.02	0.03
SO_2	8.49	9.09	9.86	8.77	9.04	9.32
NO_2	8.49	9.08	9.86	8.77	9.04	9.32
CO	8.49	9.09	9.86	8.77	9.04	9.32
O_3	8.49	9.09	9.86	8.77	9.04	9.32

4.4.3 缺失机制分析

长三角地区空气质量数据的 PM 和 IM 频率分布情况见表 4.10，每个指标中主要缺失模式为 PM，其中缺失比率最高达 63.79%。缺失区间长度从 2 天至 32 天不等。以缺失率最高的视角 AQI 为例展示缺失值的分布情况，如图 4.10 所示。图 4.10 表明 AQI 缺失区间的频率分布呈右偏。

表 4.10 点缺失(零长度)和不同长度缺失区间的频率分布

	缺失区间长度(day)	0	2	3	大于 4
AQI	频率	795	169	41	26
	比例(%)	38.39	16.32	5.94	39.35
$PM_{2.5}$	频率	866	190	41	0
	比例(%)	63.26	27.76	8.98	0
PM_{10}	频率	871	186	41	0
	比例(%)	63.76	27.23	9.01	0
SO_2	频率	860	189	41	0
	比例(%)	63.19	27.77	9.04	0
NO_2	频率	858	189	41	0
	比例(%)	63.13	27.81	9.05	0
CO	频率	860	189	41	0
	比例(%)	63.19	27.77	9.04	0
O_3	频率	857	190	41	0
	比例(%)	63.01	27.94	9.04	0

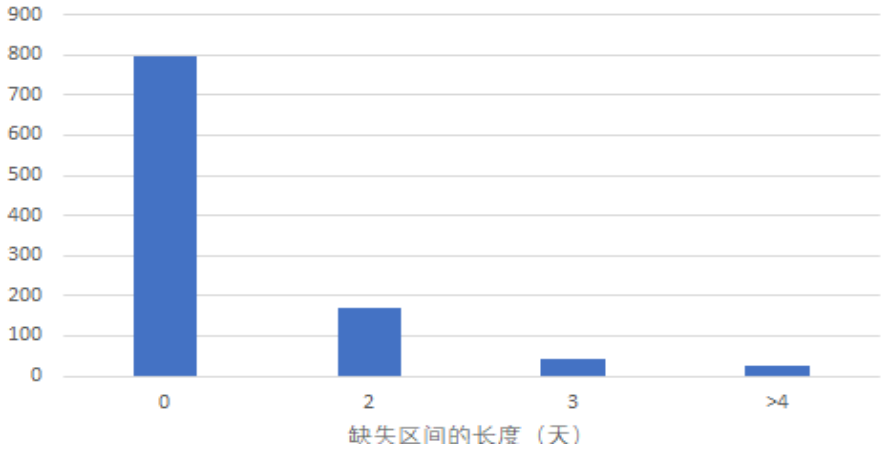


图 4.10 AQI 缺失区间的频率分布

以 2021 年的 AQI 为例对缺失数据进行可视化分析，如图 4.11 所示。图 4.11 从行和列两个维度描述缺失数据的个数，左侧数字表示各缺失情况的实例个数，

右侧数字表示有缺失值的变量(城市)个数，上方表示长三角地区的 41 个城市，下方数字表示各城市中缺失值的个数。从图中可以看出，AQI 共有 2072 个缺失值，其中安庆、黄山、六安缺失最多，为 87 个。图 4.12 左图是每个城市的缺失值数量，右图是缺失情况。左图横坐标为长三角地区各城市，纵坐标为缺失数据的个数，清晰反映缺失数据的分布情况；右图中，蓝色部分表示各城市每天的观测数据，红色部分表示缺失数据。该图不仅反映每个城市的缺失值数，还反映每个城市组合的缺失值数。

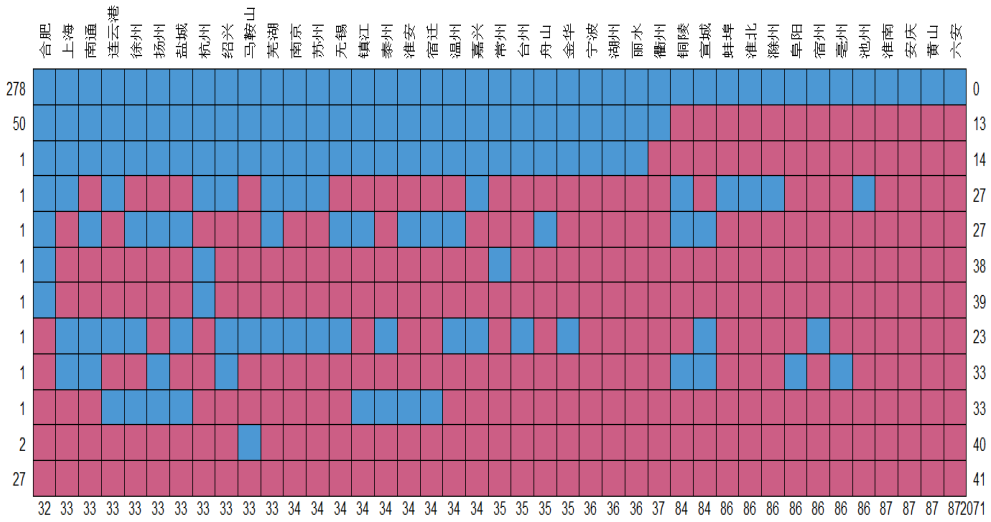


图 4.11　2021 年 AQI 缺失数据可视化

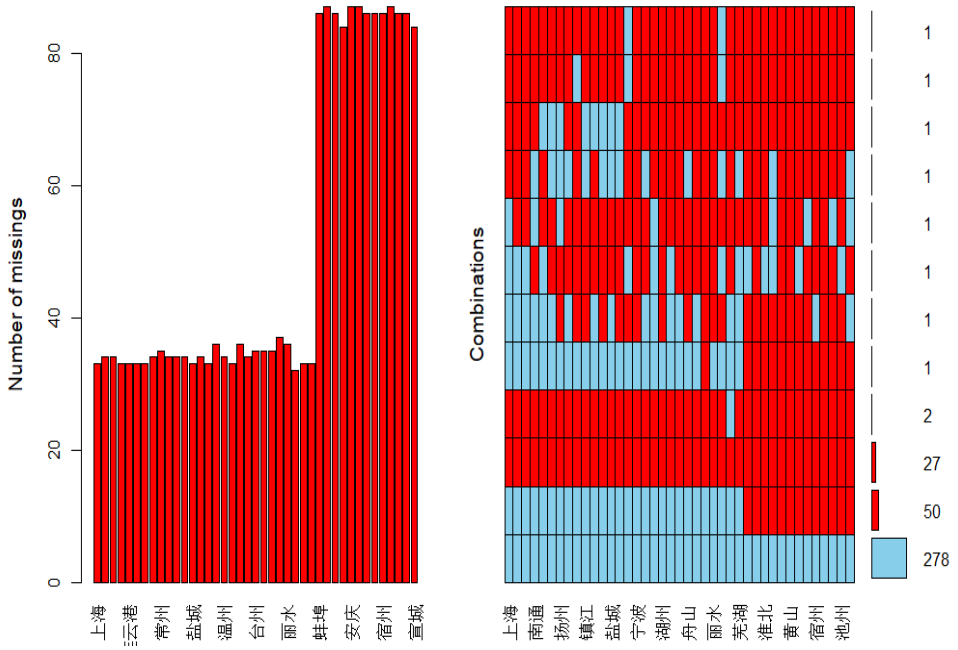


图 4.12　AQI 缺失数量及分布情况

4.4.4 有效性检验

在上述数据集的基础上，利用 GMVNFMC 方法对缺失数据进行估计，计算真实观测值与预测值的误差(RMSE、NRMSE)，并与经典函数型插补方法 SFI、HFI 以及 MVNFMC 进行比较，计算结果如表 4.11、表 4.12 所示。由表 4.11、表 4.12 可以看出，针对长三角地区多视角空气质量缺失数据集的预测插补结果，GMVNFMC 方法显著优于其他方法。

表 4.11 长三角地区空气质量各视角真实值与预测值之间的 RMSE

方法	视角 1	视角 2	视角 3	视角 4	视角 5	视角 6	视角 7
SFI	0.16	15.69	1.96	9.62	20.77	35.99	25.49
HFI	0.15	15.69	1.96	9.62	20.77	35.99	25.49
MVNFMC	0.13	20.96	2.54	124.59	117.77	66.07	16.54
GMVNFMC	**0.07**	**2.17**	**0.38**	**1.63**	**3.80**	**4.78**	**3.70**

注：粗体表示比较结果为优.

表 4.12 长三角地区空气质量各视角真实值与预测值之间的 NRMSE

方法	视角 1	视角 2	视角 3	视角 4	视角 5	视角 6	视角 7
SFI	0.25	0.50	0.28	0.35	0.32	0.61	0.47
HFI	0.25	0.50	0.28	0.35	0.32	0.61	0.47
MVNFMC	0.20	0.66	0.36	4.53	1.83	1.11	0.30
GMVNFMC	**0.12**	**0.07**	**0.05**	**0.06**	**0.06**	**0.08**	**0.07**

注：粗体表示比较结果为优.

为进一步说明插补值的有效性，以视角 AQI 为例，选取缺失率较高的 4 个城市(黄山、六安、安庆、池州)验证 GMVNFMC 方法填充值的合理性，分别计算黄山与 3 个城市之间删除缺失值后的数据、填充值的数据、填充缺失值后数据的相关系数，并绘制散点图，如图 4.13 所示。可以看出，插补值与删除缺失的数据在相关关系上一致、数值大小接近，因此，该方法估计的插补值具有一定的有效性。此外，为说明 GMVNFMC 方法插补多视角空气质量数据缺失值的准确性，对每个视角随机选取 3 个城市，分别以 3 个城市每种污染物的真实观测值和相应预测值作为横纵坐标，给出真实值关于预测值的散点图。以视角 AQI 为例，选取杭州、绍兴和铜陵绘制散点图，如图 4.14 所示(其余视角见附录图 4.15~图 4.20)。从图 4.14 可以直观地看出，每个城市的样本点大多集中在直线附近，说明真实

值与预测值之间的误差较小，即 GMVNFMC 方法的插补性能好，预测精度较高。

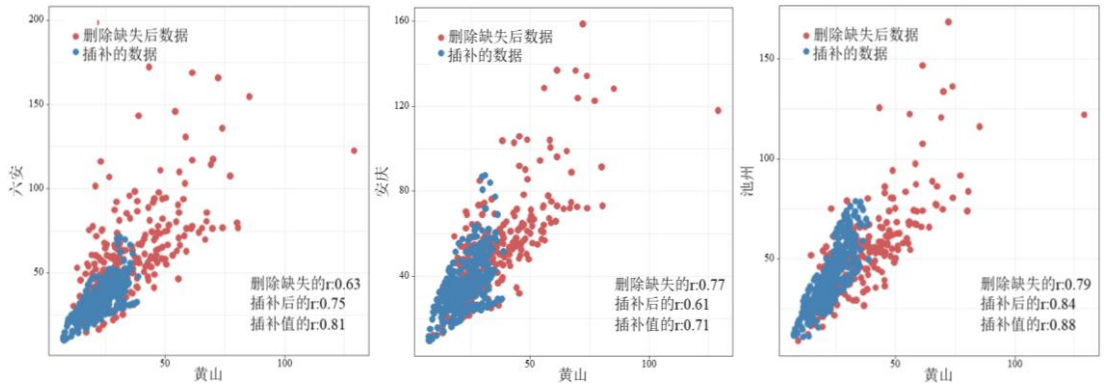


图 4.13　AQI 视角 4 个城市缺失值插补前后的相关性

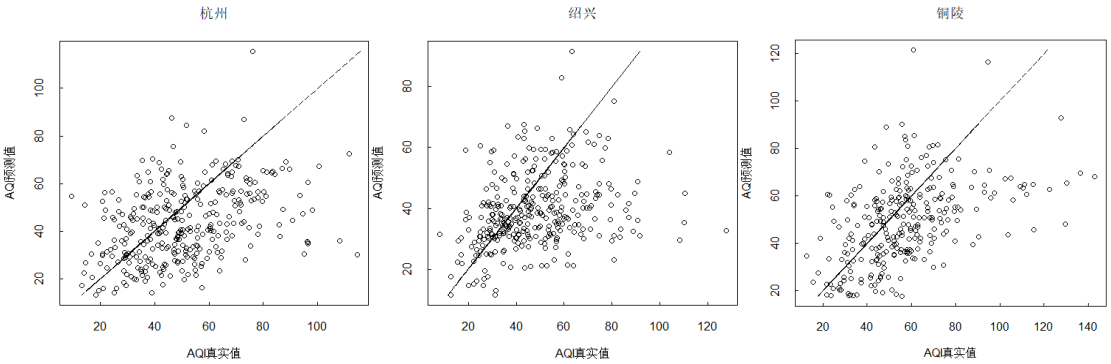


图 4.14　AQI 视角中 3 个城市真实观测值关于预测值的散点图

4.5 本章小结

本章将函数型矩阵填充方法扩展至多视角情形。首先，在缺失数据修复过程中，基于多视角学习充分综合各个视角的公共信息，通过流形学习考虑数据矩阵条目之间的非线性关系，在重构数据时基于图正则化考虑样本点之间的深层结构信息，并利用 HSIC 充分挖掘不同视角包含的互补信息，提出一种基于图正则化的多视角函数型矩阵填充方法(GMVNFMC)。其次，给出 GMVNFMC 方法的交替更新求解算法、讨论了算法的局部收敛性以及计算复杂度；最后，模拟实验结果表明，GMVNFMC 方法相较于均值填充、线性插值、KNN、SFI、HFI 以及 MVNFMC 方法不仅具有更高的插补精度，且 GMVNFMC 方法的执行时间较短，在此基础上，利用 GMVNFMC 方法对 2021 年长三角地区空气质量缺失数据进行应用，插补结果具有一定的有效性。

5 总结及展望

5.1 总结

本文首先梳理了缺失值处理技术和函数型矩阵填充方法在国内外的研究动向，总结了目前函数型矩阵填充方法存在的问题，针对性的提出了不同的解决方案。本文主要工作如下：

(1) 首先引入非负约束，提出一种融合类信息的函数型矩阵填充方法(CNFMC)，通过类信息挖掘样本曲线的相关性，并利用相似样本填充缺失值可以提高插补性能，同时应用自加权集成学习算法融合不同聚类数下的插补结果有助于进一步提高插补方法的精度。以公共交通数据集 PeMS 中车流量数据为例，验证了 CNFMC 方法的插补能力，并利用 CNFMC 方法对空气质量缺失数据进行实证应用分析，结果表明，CNFMC 方法适用场景广泛，能够保证插补的有效性和准确性，且处理时间可控。

(2) 提出基于图正则化的多视角函数型矩阵填充方法(GMVNFMC)，将单视角函数型矩阵填充方法扩展至多视角情形。GMVNFMC 方法首先基于流形学习理论在矩阵分解线性结构的基础上引入非线性关系；其次，通过最优图正则化项考虑不同视角间的深层结构信息；最后利用 HSIC 探索各个视角间的互补信息，进一步提高插补精度。该方法兼顾了视角内的深层结构信息与视角间的一致性、互补性，有效捕获数据的深层几何结构，减少了信息损失。分别在两个多视角空气质量数据集上进行模拟插补实验和实证应用分析，结果表明，相比于其他主流插补方法，GMVNFMC 方法具有更显著的插补优势。

5.2 展望

尽管本文所提出的CNFMC方法与GMVNFMC方法在数据缺失插补任务中表现出了良好的效果，但在实际应用中，数据缺失的情况复杂，且大多数据集存在异常值，在后续的工作中，我们将从如下角度展开研究：

(1) 考虑函数型数据中异常值以及噪声对缺失值插补方法性能的影响，进一步提高方法的鲁棒性。

(2) 改进CNFMC方法与GMVNFMC方法，使之能够识别不同的缺失模式，如PM缺失、IM缺失，扩大插补方法的适用范围。

(3) 受自适应算法的启发，结合机器学习，形成自适应函数型矩阵填充的深度插补思路。

(4) CNFMC方法与GMVNFMC方法具有较好的普适性，进一步可应用于其他实例场景，如气象数据、图像数据的缺失插补。

参考文献

[1]BLU T, THÉVENAZ P, UNSER M. Linear interpolation revitalized[J]. IEEE Transactions on Image Processing, 2004,13(5): 710-719.

[2]BERTSIMAS D, PAWLOWSKI C, ZHOU Y D. From predictive methods to missing dataimputation: an optimization approach[J]. The Journal of Machine Learning Research, 2017,18(1): 7133-7171.

[3]BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in neural information processing systems, 2001,14.

[4]CANDÈS E, LI X, MA Y, et al. Robust principal component analysis?[J]. Journal of the ACM (JACM), 2011,58(3): 1-37.

[5]CAI D, HE X, HAN J, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008,33(8): 1548-1560.

[6]CANDES E, RECHT B. Exact matrix completion via convex optimization[J]. Communications of the ACM, 2012,55(6): 111-119.

[7]CHEN L, CONG G, CAO X, et al. Temporal spatial-keyword top-k publish/subscribe[C]. 2015 IEEE 31St international conference on data engineering, IEEE.2015:255-266.

[8]CHIOU J M, ZHANG Y C, CHEN W H, et al. A functional data approach to missing value imputation and outlier detection for traffic flow data[J]. Transportmetrica B:Transport Dynamics, 2014,2(2): 106-129.

[9]DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of the Royal Statistical Society. Series B, Methodological, 1977,39(1): 1-38.

[10]DESCARY M H, PANARETOS V M. Functional Data Analysis by Matrix Completion[J]. The Annals of Statistics, 2019,47(1): 1-38.

[11]DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in neural

information processing systems, 2016,29.

[12] DING C, TAO L, WEI P, et al. Orthogonal nonnegative matrix tri-factorizations for clustering[C]. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA. 2006:126-135.

[13] FAN J, CHOW T W S. Sparse subspace clustering for data with missing entries and high-rank matrix completion[J]. Neural Networks, 2017,93: 36-44.

[14] FAN J, CHENG J. Matrix completion by deep matrix factorization[J]. Neural Networks, 2018,98: 34-41.

[15] FAN Y, CHEN M, PAN X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field[J]. Briefings in Bioinformatics, 2022,23(1): b361.

[16] FANG J, MENG X, QI X. A top-k POI recommendation approach based on LBSN and multi-graph fusion[J]. Neurocomputing, 2023,518: 219-230.

[17] GU Y, SONG Z, YIN J, et al. Low Rank Matrix Completion via Robust Alternating Minimization in Nearly Linear Time[J]. arXiv preprint, arXiv:2302.11068, 2023.

[18] GONG Y, LI Z, ZHANG J, et al. Missing Value Imputation for Multi-View Urban Statistical Data via Spatial Correlation Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021,35(1): 686-698.

[19] GRETTON A, BOUSQUET O, SMOLA A, et al. Measuring statistical dependence with Hilbert-Schmidt norms[C]. International conference on algorithmic learning theory, Berlin, 2005:63-77.

[20] HENRY E, KYBURG J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference by Judea Pearl[J]. Journal of Philosophy, 1991,88(8): 434-437.

[21] HAN H, HUANG M, ZHANG Y, et al. An extended-tag-induced matrix factorization technique for recommender systems[J]. Information, 2018,9(6): 143.

[22] HORVÁTH L, KOKOSZKA P. Inference for Functional Data with Applications[M]. Springer Science & Business Media, 2012.

[23] JEREZ J M, MOLINA I, GARCÍA-LAENCINA P J, et al. Missing data imputation

using statistical and machine learning methods in a real breast cancer problem[J]. Artificial Intelligence in Medicine, 2010,50(2): 105-115.

[24] JAIN P, NETRAPALLI P, SANGHAVI S. Low-rank matrix completion using alternating minimization[C]. Proceedings of the forty-fifth annual ACM symposium on Theory of computing,2013:665-674.

[25] JIAO C, GAO Y, YU N, et al. Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification[J]. IEEE journal of biomedical and health informatics, 2020,24(10): 3002-3011.

[26] JAIN S, CHOUZENOUX E, KUMAR K, et al. Graph Regularized Probabilistic Matrix Factorization for Drug-Drug Interactions Prediction[J]. IEEE Journal of Biomedical and Health Informatics, 2023,27(5): 2565-2574.

[27] JAMES G M, HASTIE T J, SUGAR C A. Principal component models for sparse functional data[J]. Biometrika, 2000,87(3): 587-602.

[28] JAMES G M. Generalized Linear Models with Functional Predictors[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2002,64(3): 411-432.

[29] JIE P, DEBASHIS P. A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components From Sparse Longitudinal Data[J]. Journal of Computational and Graphical Statistics, 2009,18(4).

[30] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009,42(8): 30-37.

[31] KIDZINSKI Ł, HASTIE T. Longitudinal Data Analysis Using Matrix Completion[J]. Stat, 2018,1050: 24.

[32] LARA-CABRERA R, GONZÁLEZ-PRIETO Á, ORTEGA F, et al. Evolving matrix-factorization-based collaborative filtering using genetic programming[J]. Applied Sciences, 2020,10(2): 675.

[33] LAIRD N M, WARE J H. Random-effects models for longitudinal data.[J]. Biometrics, 1982,38(4).

[34] LI P, CHIOU J. Functional clustering and missing value imputation of traffic flow trajectories[J]. Transportmetrica. (Abingdon, Oxfordshire, UK), 2021,9(1): 1-21.

[35] LIU M, YANG Z, LI L, et al. Auto-weighted collective matrix factorization with graph dual regularization for multi-view clustering[J]. Knowledge-Based Systems, 2023,260: 110145.

[36] LIU J, GAO J, JI S, et al. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023,204: 42-60.

[37] LIU Y, ZHENG Y, LIANG Y, et al. Urban water quality prediction based on multi-task multi-view learning[C].Proceedings of the 25th international joint conference on artificial intelligence, 2016.

[38] LEE D D, SEUNG H S. Learning the Parts of Objects by Non-Negative Matrix Factorization[J]. Nature, 1999,401(6755): 788-791.

[39] LIANG N, YANG Z, LI Z, et al. Semi-supervised multi-view clustering with graph-regularized partially shared non-negative matrix factorization[J]. Knowledge-Based Systems, 2020,190: 105185.

[40] LEE D D, SEUNG H S. Algorithms for Non-Negative Matrix Factorization[J]. Advances in neural information processing systems, 2000,13(6): 556-562.

[41] NIU D, DY J, JORDAN M. Iterative discovery of multiple alternativeclustering views[J]. IEEE transactions on pattern analysis and machine intelligence, 2013,36(7): 1340-1353.

[42] QU L, LI L, ZHANG Y, et al. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2009,10(3): 512-522.

[43] QIN M, DU Z, ZHANG F, et al. A matrix completion-based multiview learning method for imputing missing values in buoy monitoring data[J]. Information sciences, 2019,487: 18-30.

[44] RAMSAY J O, SILVERMAN B W. Functional Data Analysis[M]. New York: Springer, 2005:1-10.

[45] ROTH P. Missing data: A conceptual review for applied psychologists[J]. Personnel psychology, 1994,47(3): 537-560.

[46] RUMALING M I, CHEE F P, DAYOU J, et al. Missing Value Imputation for

PM10 Concentration in Sabah using Nearest Neighbour Method (NNM) and Expectation-Maximization (EM) Algorithm[J]. Asian Journal of Atmospheric Environment, 2020,14(1): 62-72.

[47] RAMSAY J, DALZELL C. Some Tools for Functional Data Analysis[J]. Journal of the Royal Statal Society: Series B(Mthodological), 1991,53(3): 539-561.

[48] RENNIE J, SREBRO N. Fast maximum margin matrix factorization for collaborative prediction[C]. Proceedings of the 22nd international conference on Machine learning, ACM. 2005:713-719.

[49] SHI F, DAN Z, CHEN J, et al. Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares[J]. Mathematical Problems in Engineering, 2013,2013: 1-5.

[50] SEDHAIN S, MENON A, SANNER S, et al. AutoRec: Autoencoders Meet Collaborative Filtering[C]. WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web, 2015:111-112.

[51] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]. Proceedings of the 24th international conference on world wide web, 2015:1067-1077.

[52] Van BUUREN S. Multiple imputation of discrete and continuous data by fully conditional specification[J]. Statistical methods in medical research, 2007,16(3): 219-242.

[53] WANG X, WU Y, ZHUANG D, et al. Low-rank Hankel tensor completion for traffic speed estimation[J]. IEEE Transactions on Intelligent Transportation Systems, 2023,24(5): 4862-4871.

[54] WANG C, LIU Q, WU R, et al. Confidence-aware matrix factorization for recommender systems[C]. Proceedings of the AAAI Conference on artificial intelligence, 2018,32(1).

[55] WEN Z, YIN W, ZHANG Y. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm[J]. Mathematical Programming Computation, 2012,4(4): 333-361.

[56] WANG H, NIE F, HUANG H. Multi-View Clustering and Feature Learning via

Structured Sparsity[C]. International conference on machine learning, PMLR.2013:352-360.

[57] XIE G, CHEN R, LIN Z, et al. Predicting lncRNA–disease associations based on combining selective similarity matrix fusion and bidirectional linear neighborhood label propagation[J]. Briefings in Bioinformatics, 2023,24(1): c595.

[58] YE J, ZHAO J, ZHENG F, et al. Completion and augmentation-based spatiotemporal deep learning approach for short-term metro origin-destination matrix prediction under limited observable data[J]. Neural Computing and Applications, 2023,35(4): 3325-3341.

[59] YI Y, WANG J, ZHOU W, et al. Non-negative matrix factorization with locality constrained adaptive graph[J]. IEEE Transactions on circuits and systems for videotechnology, 2019,30(8): 427-441.

[60] YAO F, MÜLLER H, WANG J. Functional linear regression analysis for longitudinal data[J]. Annals of Statistics, 2005,33(6): 2873-2903.

[61] YI X, ZHENG Y, ZHANG J, et al. ST-MVL: filling missing values in geo-sensory time series data[C]. Proceedings of the 25th International Joint Conference on Artificial Intelligence, AAAI Press. 2016.

[62] YAO F, MUELLER H, WANG J. Functional date analysis for sparse longitudinal date[J]. Journal of the American statistical association, 2005,100(470): 577-590.

[63] ZHUANG F, ZHANG Z, QIAN M, et al. Representation learning via Dual-Autoencoder for recommendation[J]. Neural Networks, 2017,90: 83-89.

[64] ZHU X, LI X, ZHANG S. Block-Row Sparse Multiview Multilabel Learning for Image Classification[J]. IEEE transactions on cybernetics, 2015,46(2): 450-461.

[65] ZHANG Z, WANG H, FAN Z, et al. Missing Road Condition Imputation Using a Multi-View Heterogeneous Graph Network From GPS Trajectory[J]. IEEE Transactions on Intelligent Transportation Systems, 2023,24(5): 4917-4931.

[66] 陈小波, 陈程, 陈蕾, 等. 基于改进低秩矩阵补全的交通量数据缺失值插补方法[J]. 交通运输工程学报, 2019,19(05): 180-190.

[67] 高海燕, 黄恒君, 王宇辰. 基于非负矩阵分解的函数型聚类算法[J]. 统计研

究, 2020,37(08): 91-103.

[68]薛娇, 傅德印, 韩海波, 等. 基于多视角学习的非负函数型矩阵填充算法[J]. 统计与决策, 2022,38(07): 5-11.

[69]张淑楠. 若干函数型混合效应模型的统计推断[D]. 浙江财经大学, 2018.

[70]张贝娜. 基于时空多视图BP神经网络的城市空气质量数据补全方法研究[J]. 浙江大学学报（理学版）, 2019,46(6): 737-744.

[71]张贤达. 矩阵分析与应用[M]. 清华大学出版社, 2013.

附录

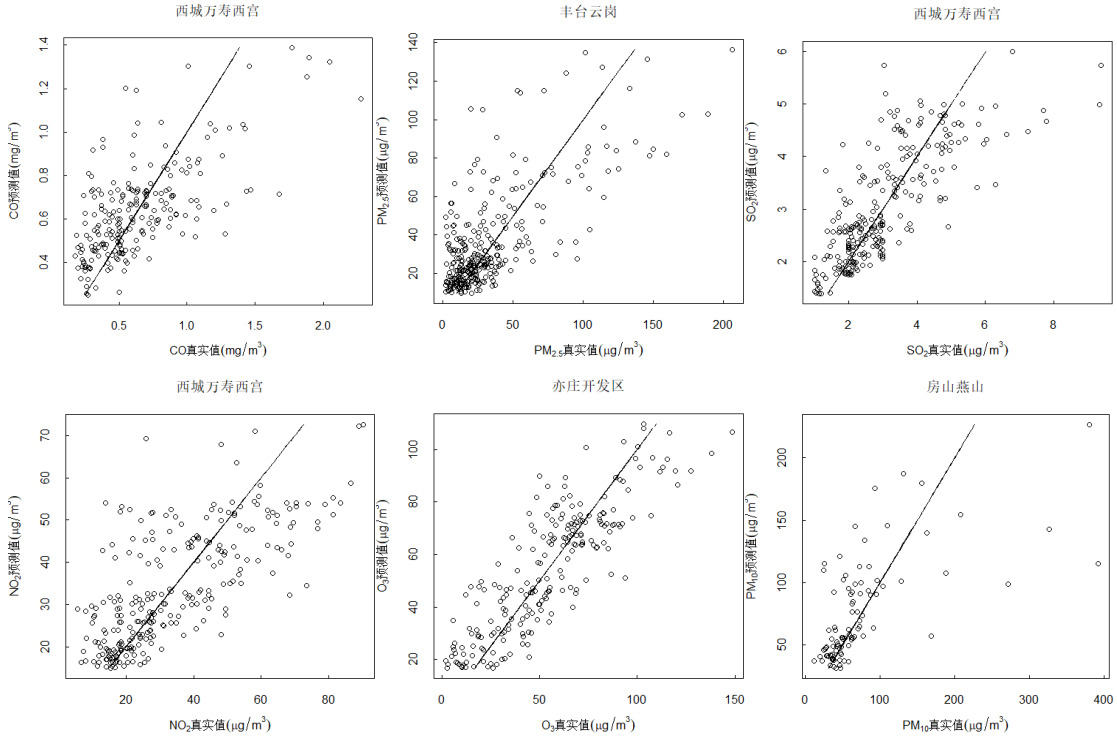


图 3.10　6 种污染物缺失率次高的站点真实观测值关于预测值的散点图

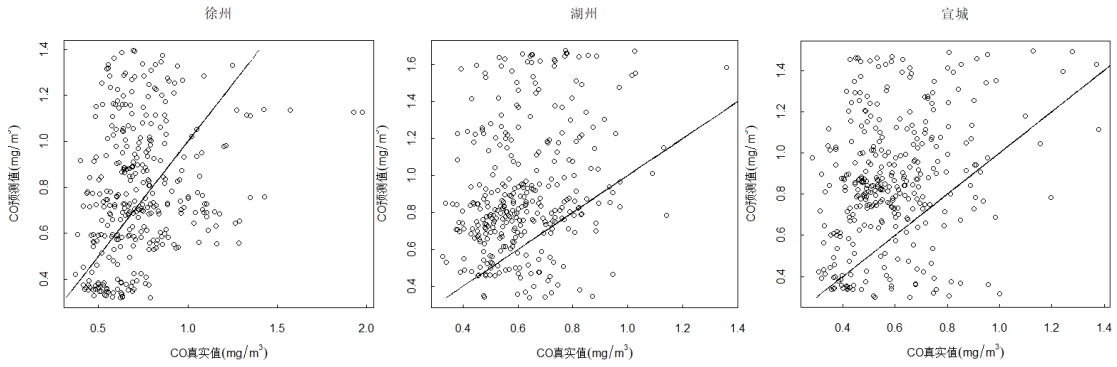


图 4.15　CO 视角中 3 个城市真实观测值关于预测值的散点图

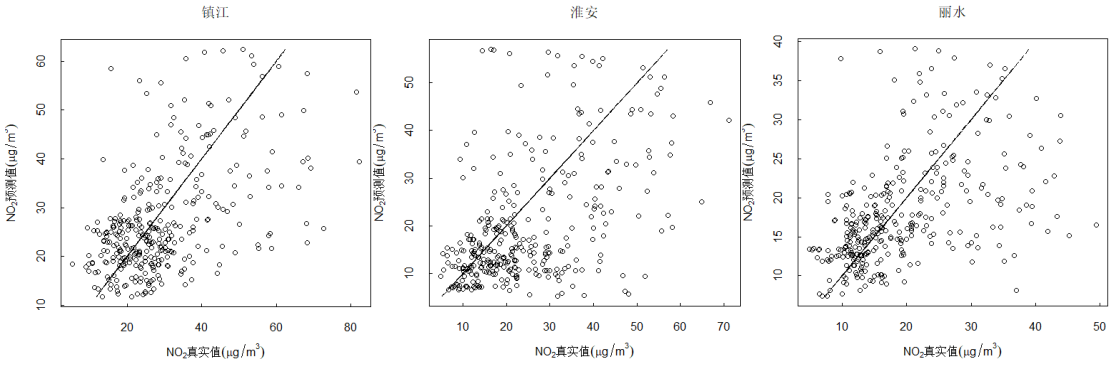


图 4.16 NO_2 视角中 3 个城市真实观测值关于预测值的散点图

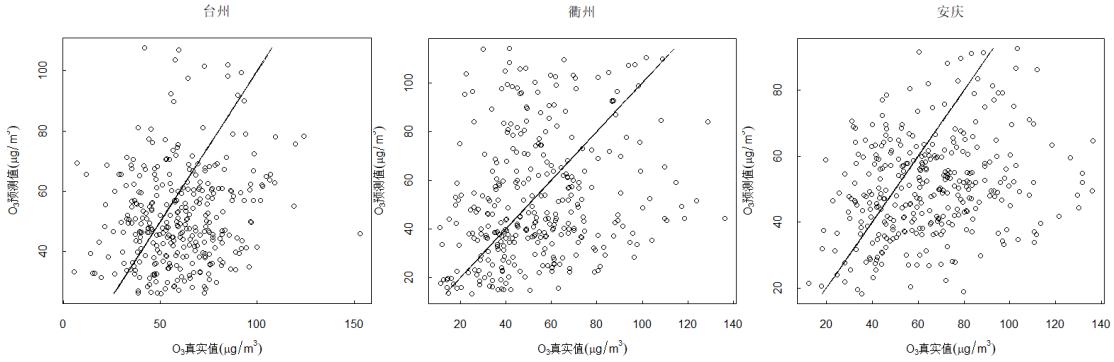


图 4.17 O_3 视角中 3 个城市真实观测值关于预测值的散点图

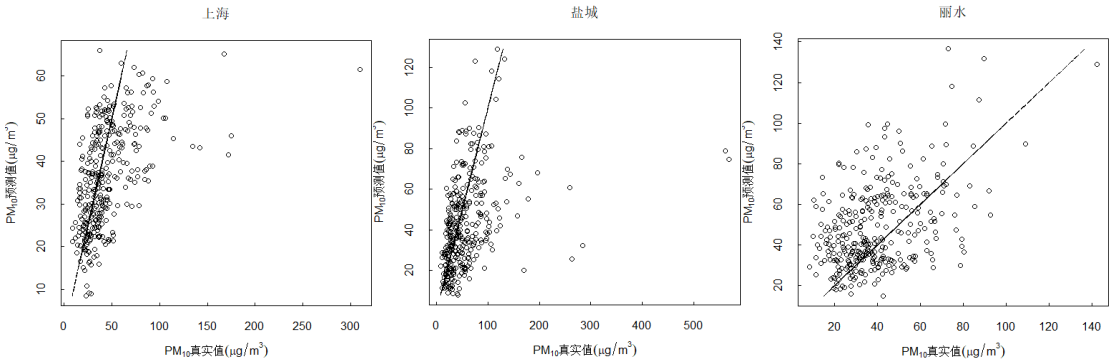


图 4.18 PM_{10} 视角中 3 个城市真实观测值关于预测值的散点图

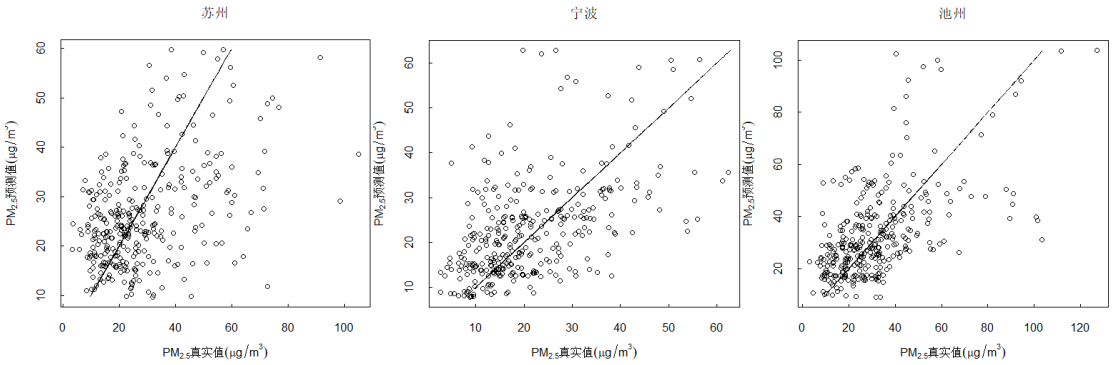


图 4.19 $PM_{2.5}$ 视角中 3 个城市真实观测值关于预测值的散点图

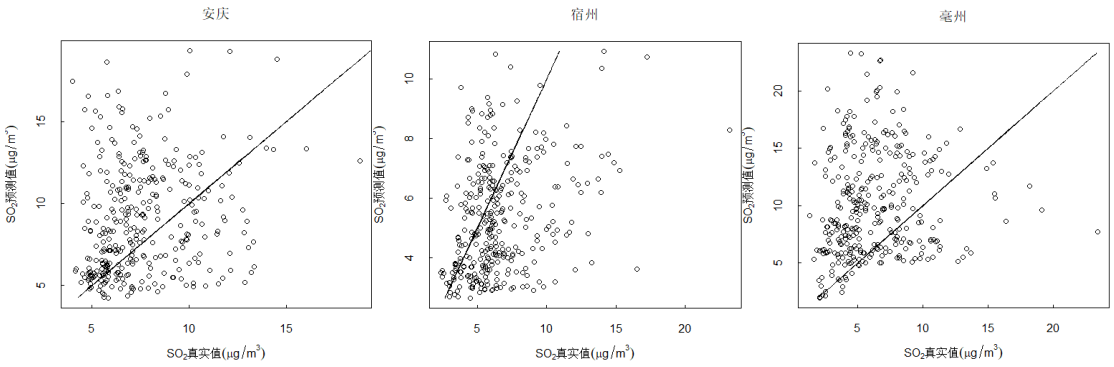


图 4.20　SO_2 视角中 3 个城市真实观测值关于预测值的散点图

➢ CNFMC 方法的 R 代码

```
#主程序
#定义以下函数的目的是实现 U、V 的更新
#由于迭代过程所用的矩阵符号和参数符号太多，在此一一做解释
  #O      投影算子，跟 Y 对齐，非 0 即 1
  #Y      带缺失的数据矩阵
  #Phi    曲线拟合的基底
  #U      初始化 U
  #V      初始化 V
  #alpha  U 的惩罚参数
  #beta   V 的惩罚参数
  #maxiter 迭代次数
NMF_SFMC<-function(O,Y,Phi,U,V,alpha,beta,maxiter)
{
  for(i in 1: maxiter)
  {
  #更新 U
  PY<-O*Y
  Phi_U<-Phi%*%U
  Phi_UV<-Phi_U%*%t(V)
  U1<-U*sqrt(((t(Phi)%*%PY)%*%V) /(t(Phi)%*%(O*Phi_UV)%*%V+alpha*U))
  #更新 V
  PY1<-t(O)*t(Y)
  Phi_U1<-Phi%*%U1
  Phi_U1V<-Phi_U1%*%t(V)
  VU1_Phi<-V %*% t(U1) %*% t(Phi)
  V1<-V*sqrt((PY1 %*% Phi_U1)/ ((t(O)*VU1_Phi)%*%Phi_U1+beta*V))
  eps<-sum((U1%*%t(V1)U%*%t(V))^2)/sum((U%*%t(V))^2)
  if(eps<-10^(7))
  {
    break
```

```
  }
  U<-U1
  V<-V1
  }
  return(list(U=U, V=V))
}
```

➢ GMVNFMC 方法的 R 代码

```
#主程序
#定义以下函数的目的是实现 U、V 的更新
#由于迭代过程所用的矩阵符号和参数符号太多，在此一一做解释
  #O    投影算子，跟 Y 同型的矩阵，若 Y 中的元素可观测，则 O 为 1，否则为 1
  #Y    带缺失的函数型数据矩阵
  #Phi   曲线拟合的基底(在 vi 可以选用样条基或者傅里基)
  #U    初始化 U_v
  #V    初始化 V_V
  #theta_v    权重参数
  #theta_vs   两个不同视角之间的权重参数
  #alpha     高低阶权重参数
  #mu       图结构的参数
  #beta      互补性的参数
  #theta_u    U_v 的拉格朗日乘子
  #theta_v     V_v 的拉格朗日乘子
  #A_v1、A_v2   各个视角的低阶、高价邻域矩阵
  #D_v1、D_v2   各个视角的低阶、高价度矩阵
  #maxiter    迭代次数
  #eps       表示两个系数矩阵之间的最大误差
  #err       表示计算模型 L2 范数误差
GMVNFMC<-function(O,Y,Phi,U,V,theta_v,theta_vs,mu,beta,theta_v1,
theta_v2,theta_v3,theta_v4,theta_v5,theta_v6,V1,V2,V3,V4,V5,V6,A,D,H)
{
  #更新 U_v
  Phi_1<-t(Phi)
  PY<-O*Y
  Phi_U<-Phi%*%U
  Phi_UV<-Phi_U%*%t(V)
  hatU<-U*sqrt((Phi_1%*% PY%*%V )/(Phi_1%*%(O*Phi_UV) %*% V))
  #更新 V_v
  PY1<-t(O)*t(Y)
  Phi_hatU<-Phi%*%hatU
  AV<-A%*%V
  V_hatU<-V %*% t(hatU)
  VhatU_Phi<-V_hatU %*% t(Phi)
```

```
  V_vs<-theta_v1*(VV1)+theta_v2*(VV2)+theta_v3*(VV3)+theta_v4*(VV4)
       +theta_v5*(VV5)+theta_v6*(VV6)
  DV<-D%*%V
  K<-H%*%V1%*%t(V1)%*%H+H%*%V2%*%t(V2)%*%H+H%*%V3%*%t(V3)%*%H+H%
    *%V4%*%t(V4)%*%H+H%*%V5%*%t(V5)%*%H+H%*%V6%*%t(V6)%*%H
  KV<-K%*%V
  hatV<-V*sqrt((theta_v*PY1%*%Phi_hatUmu*theta_v*AV)/(theta_v*(t(O)*
VhatU_Phi)%*%
       Phi_hatU + V_vs + mu*theta_v*DV + beta*theta_v*KV))
  eps<-sum((hatU%*%t(hatV)U%*%t(V))^2)/sum((U%*%t(V))^2)
  if(eps<-10^(7))
  {
    break
  }
  U<-hatU
  V<-hatV
  return(list(U=U, V=V))
}
```

攻读硕士学位期间承担的科研任务及主要成果

发表或完成的论文目录：

[1] 高海燕,**马文娟**,薛娇.融合类信息的函数型矩阵填充方法与应用[J].统计与决策,2023,39(23):40-45.

[2] 高海燕,**马文娟**,李唯欣,张悦.稀疏空气质量函数型数据插补方法实证研究[J].河北环境工程学院学报, 2023,33(05): 73-82.

[3] 高海燕,李唯欣,**马文娟**.基于缺失森林模型的稀疏函数型数据修复方法[J/OL].西华师范大学学报(自然科学版),2023,1-9.

[4] 高海燕,刘畅,**马文娟**.地表水监测缺失数据多重插补方法比较及应用[J].水文.(已录用)

科研项目目录：

(1) **主持**甘肃省优秀研究生“创新之星”项目：函数型矩阵填充方法研究及应用(2023CXZX-703)，2023.2，在研。

(2) 参与国家社会科学基金项目：大规模稀疏函数型数据修复方法与应用研究(19XTJ002)。

(3) 参与完成甘肃省优秀研究生“创新之星”项目：基于现代多重插补的稀疏函数型数据修复方法研究(2022CXZX-701)，2022.6---2023.9，已结项。

竞赛获奖：

空气质量数据的函数型聚类和缺失值插补——以北京市为例荣获第五届全国应用统计专业学位研究生案例大赛全国三等奖，2022 年 8 月。

致谢

时光荏苒，岁月如梭，转瞬间，我即将完成我的研究生学业。在这段求知路上，我得到了许多人的帮助和支持，在此，我怀着无比感激的心情，向那些给予我关爱、支持和帮助的人们致以最诚挚的谢意。

感谢我的导师高海燕教授。您是我求学道路上的指引者和启迪者，从我进入研究生阶段开始，您就给予了我无私的指导和悉心的教诲。在学术论文的撰写和研究项目的开展中，您总是耐心倾听我的问题和困惑，并给予我宝贵的建议和指引。您的严谨治学态度和渊博的学识使我受益匪浅，让我懂得了追求卓越的重要性。感谢您对我的信任和支持，您一直是我学习和人生的榜样。

感谢我的家人。感谢父母多年来给予我的支持和鼓励，让我在学业和生活中始终坚定前行。您们的无私奉献与关爱是我前进的动力，在这个充满挑战的学术道路上，你们一直给予我坚定的支持。

感谢高家班的每位同学。在这个团结友爱的科研团队中，每一个人都给予了我热情的帮助和支持。无论是开展模拟实验、解决难题还是学术交流，大家总是密切合作，共同进步。我们一起度过了忙碌而充实的时光，留下了难忘的回忆，衷心祝愿每位团队成员未来的道路越来越宽广，前途更加光明。

感谢我的朋友们。在学习和生活中，真挚的友谊让我更加坚定和自信，也能在困难时给予我无限的鼓励和支持。因为有你们的陪伴，我才能度过这段旅程，继续开启新的人生篇章。

此外，我要感谢那些无法一一列举名字的人。感谢你们在我生命中默默出现，或者短暂相遇，或是默默支持，甚至只是一次鼓励的微笑，都让我深刻感受到了温暖和力量。谢谢你们。总之，在这段学习旅途中，有诸多的人和事值得我去感恩。我也将会怀着感恩之心，继续前行，努力成为自己、家人和朋友们的骄傲。

山重水复疑无路，柳暗花明又一村。千山万水，皆为知识之途。愿与诸君共勉，共创美好未来。